#10088895 – 2 Fast & Focus

STIC Search Report
EIC 2100

STIC Database Tracking Number: 163582

TO: Gwen Liang **Location: RND 3B11** **Art Unit : 2162** **Wednesday, August 24, 2005** **Case Serial Number: 10/088895**	**From: Carol Wong** **Location: EIC 2100** **RND 4A30** **Phone: 272-3513** **carol.wong@uspto.gov**

Search Notes

Dear Examiner Liang,

Attached are the search results (from commercial databases) for your case.

Color tags mark the patents/articles which appear to be most relevant to the case.

Please call if you have any questions or suggestions for additional terminology, or a different approach to searching the case.

Thanks,
Carol

#10088895 - 2

Access DB# _____

Fast & Focus **SEARCH REQUEST FORM**

Scientific and Technical Information Center

Requester's Full Name: Gwen Liang Examiner #: 79180 Date: 8-24-05
Art Unit: 2162 Phone Number ~~30~~ X24038 Serial Number: 10/088,895
Mail Box and Bldg/Room Location: RND 3B-11 Results Format Preferred (circle): PAPER DISK E-MAIL

If more than one search is submitted, please prioritize searches in order of need.

****

Please provide a detailed statement of the search topic, and describe as specifically as possible the subject matter to be searched. Include the elected species or structures, keywords, synonyms, acronyms, and registry numbers, and combine with the concept or utility of the invention. Define any terms that may have a special meaning. Give examples or relevant citations, authors, etc, if known. Please attach a copy of the cover sheet, pertinent claims, and abstract.

Title of Invention: Method of Thematic Classification of Documents - - -

Inventors (please provide full names): BIETTRON, Laurent; PALLU, Frederic; TRICOT, Sylvie

Earliest Priority Filing Date: 9-24-99 * Assignee = France Telecom

***For Sequence Searches Only* Please include all pertinent information (parent, child, divisional, or issued patent numbers) along with the appropriate serial number.**

Inventor Search "Gerard Salton" related to document classification (or categorization) - particularly the method steps of assigning coefficient to show the relevance between an element (term) and a theme, as claimed in claim 14, steps 14-10, 14-11, 14-12.

****

STAFF USE ONLY **Type of Search** **Vendors and cost where applicable**

AMENDMENTS TO THE CLAIMS:

This listing of claims will replace all prior versions, and listings, of claims in the application:

LISTING OF CLAIMS:

1-13. (canceled)

14. (currently amended) A method of thematically classifying documents, in particular for making up or updating thematic databases for a search engine, the method comprising the following steps:

14-1 - manually and/or automatically selecting a sample of documents representative of each theme;

14-2 - automatically identifying within the selected documents elements that are characteristic of each theme;

14-3 - automatically allocating a coefficient to each identified element, which coefficient is representative of the relevance of said element relative to the corresponding theme;

14-4 - downloading documents from a computer network;

14-5 - for each downloaded document to be classified, identifying said theme-characterizing elements that are contained in the document for each of the themes, and for each theme corresponding to the elements, using the coefficients allocated to said elements to calculate [[the]] a characteristic value ~~of a characteristic~~ representative of the relevance of that theme for the document, in order to decide whether or not the document

relates to the theme, said theme-characterizing elements identification and calculation steps being performed automatically for each document downloaded from [[a]] the computer network;

14-6 - automatically classifying the downloaded documents as a function of the themes with which they deal; [[and]]

14-7 - automatically storing the documents classified thematically in databases that can be interrogated on the basis of themes contained in a request; and

14-8 - making the databases available to users who interrogate the databases on the basis of themes contained in a request;

14-9 and the step of allocating said coefficient to each identified element comprises the following steps for each theme:

* 14-10 - automatically calculating [[the]] a frequency of the element in the selected documents relating to the theme;

* 14-11 - automatically calculating [[the]] a frequency of the element in the selected documents that do not relate to the theme; and

* 14-12 - automatically calculating the ratio of the calculated frequencies.

15. (previously presented) A method according to claim 14, further comprising the step of automatically sorting themes in a theme tree structure in decreasing order of coefficients.

Example= #10/0.88.98_2

8-23-05
Went thru with Lulee

Categories = travel, car, clothing
term = food

5 documents = food

			key
doc#1	(travel)	5	(food)
doc#2	(travel)	15	(food)
doc#3	(car)	2	(food)
doc#4	(computer)	1	(food)
doc#5	(clothing)	2	(food)

Coefficient of (food) related to (travel)

$$oe = \frac{\text{freq of food in travel} \quad (5+15)/2 = 10}{\text{freq of food in others} \quad 5/3 = \frac{5}{3}} = 6$$

To search: Gerard Salton (inventor)

File 348:EUROPEAN PATENTS 1978-2005/Aug W02
(c) 2005 European Patent Office
File 349:PCT FULLTEXT 1979-2005/UB=20050818,UT=20050811
(c) 2005 WIPO/Univentio
File 324:German Patents Fulltext 1967-200532
(c) 2005 Univentio

Set	Items	Description
S1	13	GERALD(1W)SALTON
S2	0	AU='SALTON G'
S3	432430	CLASSIFY? OR CLASSIFIE? ? OR CLASSIFICAT? OR CATEGORI? OR -CATEGORY?
S4	42687	CATALOG? OR TAXONOM? OR AUTOCLASSIF? OR AUTOCATEGOR? OR AUTOCATALOG?
S5	10	S1 AND S3:S4
S6	10	IDPAT (sorted in duplicate/non-duplicate order)
S7	10	IDPAT (primary/non-duplicate records only)
S8	0	AU=SALTON G?
S9	32	SALTON(1N)(GERALD OR GERARD)
S10	27	S9 AND (S3:S4 OR FREQUEN?)
S11	17	S10 NOT S7
S12	17	IDPAT (sorted in duplicate/non-duplicate order)
S13	17	IDPAT (primary/non-duplicate records only)
S14	15	S13 AND AC=US/PR
S15	12	S14 AND AY=(1970:1999)/PR
S16	12	S13 AND PY=1970:1999
S17	13	S15:S16

? t7/5,k/all

7/5,K/1 (Item 1 from file: 348)
DIALOG(R)File 348:EUROPEAN PATENTS

01452298
Method of categorizing **a document into a document hierarchy**
Methode zur Einordnung eines Dokuments in einen Dokumentenbestand
Procede de categorisation **d'un document dans une hierarchie de documents**
PATENT ASSIGNEE:
Siemens Business Services GmbH & Co. OHG, (2760190), Otto-Hahn-Ring 6, 81739 Munchen, (DE), (Applicant designated States: all)
INVENTOR:
Hofer, Hardy, Rochusweg 52, 33102 Paderborn, (DE)
LEGAL REPRESENTATIVE:
Berg, Peter et al (89732), European Patent Attorney, Siemens AG, Postfach 22 16 34, 80506 Munchen, (DE)
PATENT (CC, No, Kind, Date): EP 1244027 A1 020925 (Basic)
APPLICATION (CC, No, Date): EP 2001107285 010323;
DESIGNATED STATES: AT; BE; CH; CY; DE; DK; ES; FI; FR; GB; GR; IE; IT; LI; LU; MC; NL; PT; SE; TR
EXTENDED DESIGNATED STATES: AL; LT; LV; MK; RO; SI
INTERNATIONAL PATENT CLASS: G06F-017/30

ABSTRACT EP 1244027 A1 (Translated)
Method for coordinating a new document in an existing data stock structured by **classification** criteria fixes each closest document in relation to the new document with a minimum clearance according to a preset selection fun
Each closest document is fixed in relation to a new document and has a minimum clearance from the new document in respect of a preset clearance measurement according to a preset selection function. **Classification** criteria for the new document are deduced from the **classification** criteria of the closest document and represent a tree structure.
TRANSLATED ABSTRACT WORD COUNT: 89

ABSTRACT EP 1244027 A1
Methode zum Einordnen eines neuen Dokuments in einen bestehenden Datenbestand, der durch Ordnungskriterien strukturiert ist, wobei zu dem neuen Dokument dasjenige nachstliegende Dokument bestimmt wird, das zu dem neuen Dokument bezuglich eines vorgegebenen Abstandsmases einen minimalen Abstand hat, und die Ordnungskriterien des neuen Dokuments aus den Ordnungskriterien des nachstliegenden Dokuments abgeleitet werden.
ABSTRACT WORD COUNT: 53

LEGAL STATUS (Type, Pub Date, Kind, Text):
Application: 020925 A1 Published application with search report
Examination: 030521 A1 Date of request for examination: 20030325
Examination: 050105 A1 Date of dispatch of the first examination report: 20041122
LANGUAGE (Publication,Procedural,Application): German; German; German
FULLTEXT AVAILABILITY:

Available Text	Language	Update	Word Count
CLAIMS A	(German)	200239	178
SPEC A	(German)	200239	828
Total word count - document A			1006
Total word count - document B			0
Total word count - documents A + B			1006

Method of categorizing **a document into a document hierarchy**

Procede de categorisation **d'un document dans une hierarchie de documents**

...ABSTRACT Translated)
Method for coordinating a new document in an existing data stock structured by **classification** criteria fixes each closest document in relation to the new document with a minimum clearance...

...new document in respect of a preset clearance measurement according to a preset selection function. **Classification** criteria for the new document are deduced from the **classification** criteria of the closest document and represent a tree structure.

...SPECIFICATION dem Vektorraum-Modell verwendet. Dieses ist z.B. in "Introduction to Modern Information Retrieval" von **Gerald Salton** , McGraw Hill 1983, S.121-122, beschrieben. Eine Ubersicht hieruber findet sich auch in der...

7/5,K/2 (Item 2 from file: 348)
DIALOG(R)File 348:EUROPEAN PATENTS

00843586
Automatic method of identifying sentence boundaries in a document image
Automatisches Verfahren zum Identifizieren von Satzgrenzen in der Abbildung eines Dokumentes
Procede automatique d'identification des limites de phrases dans une image d'un document
PATENT ASSIGNEE:
XEROX CORPORATION, (219783), Xerox Square, Rochester New York 14644, (US) , (applicant designated states: DE;FR;GB)
INVENTOR:
Bloomberg, Dan S., 1013 Paradise Way, Palo Alto, California 94306, (US)
LEGAL REPRESENTATIVE:
Grunecker, Kinkeldey, Stockmair & Schwanhausser Anwaltssozietat (100721) , Maximilianstrasse 58, 80538 Munchen, (DE)
PATENT (CC, No, Kind, Date): EP 779594 A2 970618 (Basic)
EP 779594 A3 980114
APPLICATION (CC, No, Date): EP 96308997 961211;
PRIORITY (CC, No, Date): US 572597 951214
DESIGNATED STATES: DE; FR; GB
INTERNATIONAL PATENT CLASS: G06K-009/32;
ABSTRACT EP 779594 A2
A method of automatically identifying sentence boundaries in a document image without performing OCR. The identification process begins by selecting a connected component from the multiplicity of connected components of a text line. Next, it is determined whether the selected connected component might represent a period based upon its shape. If the selected connected component is dot shaped, then it is determined whether the selected connected component might represent a colon. Finally, if the selected connected component is dot shaped and not part of a colon, the selected connected component is labeled as a sentence boundary.
ABSTRACT WORD COUNT: 97

LEGAL STATUS (Type, Pub Date, Kind, Text):
Withdrawal: 050112 A2 Date application deemed withdrawn: 20040701
Application: 970618 A2 Published application (Alwith Search Report ;A2without Search Report)
Search Report: 980114 A3 Separate publication of the European or International search report

Change: 980617 A2 Representative (change)
Examination: 980909 A2 Date of filing of request for examination: 980714
Examination: 991027 A2 Date of dispatch of the first examination report: 19990910
LANGUAGE (Publication,Procedural,Application): English; English; English
FULLTEXT AVAILABILITY:

Available Text	Language	Update	Word Count
CLAIMS A	(English)	EPAB97	765
SPEC A	(English)	EPAB97	11527
Total word count - document A			12292
Total word count - document B			0
Total word count - documents A + B			12292

...SPECIFICATION is not limited.
Quantitative content analysis relies upon statistical properties of text to produce summaries. **Gerald Salton** discusses the use of quantitative content analysis to summarize documents in "Automatic Text Processing" (1989...significantly larger or smaller than the median font size for the document. Preferably, processor 11 **classifies** a block as non-conforming if its median height varies more than 15% from the...very frequently in natural language text. Most pronouns, prepositions, determiners, and "to be" verbs are **classified** as drop words. Thus, for example, words such as "and, a, the, on, by, about...

7/5,K/3 (Item 3 from file: 348)
DIALOG(R)File 348:EUROPEAN PATENTS

00843585
Automatic method of identifying drop words in a document image without performing OCR
Automatisches Verfahren zum Identifizieren von Wegfallwortern in der Abbildung eines Dokumentes ohne Verwendung vom OCR
Procede automatique d'identification des mots a omettre dans une image d'un document, sans la mise en oeuvre d'OCR
PATENT ASSIGNEE:
XEROX CORPORATION, (219783), Xerox Square, Rochester, New York 14644, (US), (Proprietor designated states: all)
INVENTOR:
Chen, Francine R., 975 Sherman Avenue, Menlo Park, California 94025, (US)
Tukey, John W., 115 Arreton Road, PO Box 2043, Princeton, New Jersey 08540-2043, (US)
LEGAL REPRESENTATIVE:
Grunecker, Kinkeldey, Stockmair & Schwanhausser Anwaltssozietat (100721), Maximilianstrasse 58, 80538 Munchen, (DE)
PATENT (CC, No, Kind, Date): EP 779592 A2 970618 (Basic)
EP 779592 A3 980114
EP 779592 B1 011024
APPLICATION (CC, No, Date): EP 96308996 961211;
PRIORITY (CC, No, Date): US 572847 951214
DESIGNATED STATES: DE; FR; GB
INTERNATIONAL PATENT CLASS: G06K-009/42; G06T-007/00; G06F-017/30
CITED PATENTS (EP B): EP 544432 A

ABSTRACT EP 779592 A2
A method of automatically identifying drop words in a document image without performing OCR. First, the document image is analyzed to identify word equivalence classes, each of which represents at least one word of

the multiplicity of words included in the document. Second, for each word equivalence class, the likelihood that it is not a drop word is determined. Third, document length is analyzed to determine whether the document is short. For a short document, the number of word equivalence classes identified as drop words based upon their likelihood is proportional to document length. For long documents, a fixed number of word equivalence classes are identified as drop words based upon the likelihood that they are not drop words.

ABSTRACT WORD COUNT: 120

NOTE:

Figure number on first page: 11

LEGAL STATUS (Type, Pub Date, Kind, Text):

Grant:	011024	B1	Granted patent
Application:	970618	A2	Published application (A1with Search Report ;A2without Search Report)
Oppn None:	021016	B1	No opposition filed: 20020725
Change:	980107	A2	International patent **classification** (change)
Search Report:	980114	A3	Separate publication of the European or International search report
Change:	980617	A2	Representative (change)
Examination:	980909	A2	Date of filing of request for examination: 980714
Examination:	990825	A2	Date of dispatch of the first examination report: 19990709

LANGUAGE (Publication,Procedural,Application): English; English; English

FULLTEXT AVAILABILITY:

Available Text	Language	Update	Word Count
CLAIMS A	(English)	EPAB97	597
CLAIMS B	(English)	200143	656
CLAIMS B	(German)	200143	654
CLAIMS B	(French)	200143	832
SPEC A	(English)	EPAB97	11496
SPEC B	(English)	200143	11370
Total word count - document A			12095
Total word count - document B			13512
Total word count - documents A + B			25607

LEGAL STATUS (Type, Pub Date, Kind, Text):

...International patent **classification** (change)

Search Report...

...SPECIFICATION is not limited.

Quantitative content analysis relies upon statistical properties of text to produce summaries. **Gerald Salton** discusses the use of quantitative content analysis to summarize documents in "Automatic Text Processing" (1989...significantly larger or smaller than the median font size for the document. Preferably, processor 11 **classifies** a block as non-conforming if its median height varies more than 15% from the...very frequently in natural language text. Most pronouns, prepositions, determiners, and "to be" verbs are **classified** as drop words. Thus, for example, words such as "and, a, the, on, by, about...

...SPECIFICATION is not limited.

Quantitative content analysis relies upon statistical properties of text to produce summaries. **Gerald Salton** discusses the use of quantitative content analysis to summarise documents in "Automatic Text Processing" (1989...significantly larger or smaller than the median font size for the document. Preferably, processor 11 **classifies** a block as non-conforming if its median height varies more than 15% from the...very frequently in natural language text. Most pronouns, prepositions,

determiners, and "to be" verbs are **classified** as drop words. Thus, for example, words such as "and, a, the, on, by, about...

7/5,K/4 (Item 4 from file: 348)
DIALOG(R)File 348:EUROPEAN PATENTS

00791575
Automatic method of generating thematic summaries
Automatisches Verfahren zur Erzeugung von thematischen Zusammenfassungen
Methode automatique pour la generation de resumes thematiques
PATENT ASSIGNEE:
XEROX CORPORATION, (219783), Xerox Square, Rochester, New York 14644, (US), (Proprietor designated states: all)
INVENTOR:
Chen, Francine R., 975 Sherman Avenue, Menlo Park, CA 94025, (US)
LEGAL REPRESENTATIVE:
Grunecker, Kinkeldey, Stockmair & Schwanhausser Anwaltssozietat (100721), Maximilianstrasse 58, 80538 Munchen, (DE)
PATENT (CC, No, Kind, Date): EP 737927 A2 961016 (Basic)
EP 737927 A3 981118
EP 737927 B1 011205
APPLICATION (CC, No, Date): EP 96302250 960329;
PRIORITY (CC, No, Date): US 422573 950414
DESIGNATED STATES: DE; FR; GB
INTERNATIONAL PATENT CLASS: G06F-017/27; G06F-017/30
CITED PATENTS (EP B): US 5384703 A
CITED REFERENCES (EP B):
EDMUNDSON, H.P.: "New methods in automatic extracting" JOURNAL OF THE ASSOCIATION FOR COMPUTING MACHINERY, vol. 16, April 1969, pages 264-285, XP002078269
"METHOD FOR AUTOMATIC EXTRACTION OF RELEVANT SENTENCES FROM TEXTS" IBM TECHNICAL DISCLOSURE BULLETIN, vol. 33, no. 6A, November 1990, page 338/339 XP002015802
BLACK W J ET AL: "A PRACTICAL EVALUATION OF TWO RULE-BASED AUTOMATIC ABSTRACTING TECHNIQUES" EXPERT SYSTEMS FOR INFORMATION MANAGEMENT, vol. 1, no. 3, 1988, pages 159-177, XP002015761
LUHN, H.P.: "The automatic creation of literature abstracts" IBM JOURNAL,April 1958, page 159-165 XP002078270;

ABSTRACT EP 737927 A2
A technique for automatically generating thematic summaries for machine readable representations of documents. The technique begins with identification of thematic terms (42,44) within the document. Afterward, each sentence of the document is scored (46-58) based upon the number of thematic terms contained within the sentence. The highest scoring sentences are selected (62) as thematic sentences. (see image in original document)
ABSTRACT WORD COUNT: 72
NOTE:
Figure number on first page: 2

LEGAL STATUS (Type, Pub Date, Kind, Text):

Examination:	000621 A2	Date of dispatch of the first examination report: 20000504
Application:	961016 A2	Published application (A1with Search Report ;A2without Search Report)
Oppn None:	021127 B1	No opposition filed: 20020906
Grant:	011205 B1	Granted patent
Change:	980617 A2	Representative (change)

Search Report: 981118 A3 Separate publication of the European or International search report
Examination: 990714 A2 Date of filing of request for examination: 990518
LANGUAGE (Publication,Procedural,Application): English; English; English
FULLTEXT AVAILABILITY:

Available Text	Language	Update	Word Count
CLAIMS A	(English)	EPAB96	323
CLAIMS B	(English)	200149	348
CLAIMS B	(German)	200149	352
CLAIMS B	(French)	200149	402
SPEC A	(English)	EPAB96	2349
SPEC B	(English)	200149	3388
Total word count - document A			2672
Total word count - document B			4490
Total word count - documents A + B			7162

...SPECIFICATION is not limited.
Quantitative content analysis relies upon statistical properties of text to produce summaries. **Gerald Salton** discusses the use of quantitative content analysis to summarize documents in "Automatic Text Processing" (1989...very frequently in natural language text. Most pronouns, prepositions, determiners, and "to be" verbs are **classified** as stop words. Thus, for example, words such as "and, a, the, on, by, about...

...SPECIFICATION is not limited.
Quantitative content analysis relies upon statistical properties of text to produce summaries. **Gerald Salton** discusses the use of quantitative content analysis to summarize documents in "Automatic Text Processing" (1989...very frequently in natural language text. Most pronouns, prepositions, determiners, and "to be" verbs are **classified** as stop words. Thus, for example, words such as "and, a, the,

7/5,K/5 (Item 5 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT

00912809 **Image available**
SYSTEM FOR FULFILLING AN INFORMATION NEED USING EXTENDED MATCHING TECHNIQUES
SYSTEME PERMETTANT DE REPONDRE A UN BESOIN D'INFORMATION PAR DES TECHNIQUES D'APPARIEMENT APPROFONDIES
Patent Applicant/Assignee:
GLOBAL INFORMATION RESEARCH AND TECHNOLOGIES LLC, 236 Huntington Avenue, Boston, MA 02115-4701, US, US (Residence), US (Nationality)
Inventor(s):
SCHABES Yves, c/o Teragram, 236 Huntington Avenue, Boston, MA 02115, US,
ROCHE Emmanuel, c/o Teragram, 236 Huntington Avenue, Boston, MA 02115, US
,
Legal Representative:
HAMILTON John A (agent), Choate, Hall & Stewart, Exchange Place, 53 State Street, Boston, MA 02109, US,
Patent and Priority Information (Country, Number, Date):
Patent: WO 200246970 A2-A3 20020613 (WO 0246970)
Application: WO 2001US46542 20011205 (PCT/WO US01046542)
Priority Application: US 2000251608 20001205
Designated States:
(Protection type is "patent" unless otherwise stated - for applications

prior to 2004)
AE AG AL AM AT AU AZ BA BB BG BR BY BZ CA CH CN CO CR CU CZ DE DK DM DZ
EC EE ES FI GB GD GE GH GM HR HU ID IL IN IS JP KE KG KP KR KZ LC LK LR
LS LT LU LV MA MD MG MK MN MW MX MZ NO NZ PL PT RO RU SD SE SG SI SK SL
TJ TM TR TT TZ UA UG UZ VN YU ZA ZW
(EP) AT BE CH CY DE DK ES FI FR GB GR IE IT LU MC NL PT SE TR
(OA) BF BJ CF CG CI CM GA GN GQ GW ML MR NE SN TD TG
(AP) GH GM KE LS MW MZ SD SL SZ TZ UG ZM ZW
(EA) AM AZ BY KG KZ MD RU TJ TM
Main International Patent Class: G06F-017/30
Publication Language: English
Filing Language: English
Fulltext Availability:
Detailed Description
Claims
Fulltext Word Count: 24863

English Abstract
The invention offers new approaches to fulfilling an information need, in particular to finding a result for a query based on a large body of information such as a collection of documents. The invention accepts a query containing an unspecified portion that expresses the information need. The invention locates matches for the query within a body of information and returns the matches or portions thereof in addition to or instead of identifiers for documents in which the matches are found. The invention allows placement of term ordering restrictions, and allows intervening words between the search terms as they appear in the searched documents or contexts. The invention ranks the matches in order to provide the most relevant information. One preferred method of ranking considers the number of instances of a match among a plurality of documents. The invention further defines a new type of index that includes contexts in which terms occur and provides methods of searching such indices to fulfill an information need.

French Abstract
L'invention presente de nouvelles approches permettant de repondre a un besoin d'information, en particulier de trouver un resultat a une interrogation en fonction d'un grand nombre d'informations, tel qu'une collection de documents. Selon l'invention, le systeme accepte une interrogation contenant une partie non specifiee qui exprime le besoin d'information. Ce systeme localise des correspondances pour cette interrogation dans un corps d'informations, et renvoie ces correspondances, ou des parties de celles-ci, en plus ou a la place d'identificateurs de documents dans lesquels on trouve ces correspondances. L'invention permet de placer des restrictions de classement de termes, et elle permet de faire intervenir des mots entre les termes de recherche, a mesure qu'ils apparaissent dans les documents ou les contextes explores. L'invention classe les correspondances dans l'ordre pour fournir l'information la plus pertinente. Dans un procede de classement prefere, le nombre d'exemples de correspondance parmi une pluralite de documents est pris en consideration. Le systeme definit egalement un nouveau type d'index qui comporte des contextes dans lesquels des termes apparaissent, et met a disposition des procedes de recherche de tels indices pour repondre a un besoin d'information.

Legal Status (Type, Date, Text)
Publication 20020613 A2 Without international search report and to be republished upon receipt of that report.
Examination 20030206 Request for preliminary examination prior to end of 19th month from priority date
Search Rpt 20040226 Late publication of international search report

Republication 20040226 A3 With international search report.

Fulltext Availability:
Detailed Description
Claims

Detailed Description
... portion is only partially unspecified, in which case it includes a restriction that defines a **category** of terms that are acceptable matches for the unspecified portion. The restriction may indicate that an unspecified portion of a query must meet a **category** criterion, a morphological or syntactic criterion, or a criterion defined by a computer program. For...26 and 27. In certain embodiments of the invention data stored in memory 15 includes **category** lists 28, dictionaries 29, and index or indices 50 which are discussed farther below.

Applications...can also be stored on peripheral storage 43. Data stored in memory 43 preferably includes **category** lists 48 and dictionaries 49. In preferred embodiments of the invention these are identical to **category** lists 28 and dictionaries 29 stored on query server(s) 4. In certain embodiments of the invention query servers 4 and indexing computer 6 access the same **category** lists and dictionaries. Index or indices 50 generated by indexing computer 6 are also preferably...

...upon the indexer described in "The SMART Retrieval System: Experiments in Automatic Document Processing" by **Gerald Salton** (Prentice-Hall, Inc. (1971)) and "A Theory of Indexing" also by **Gerald Salton** (J. W. Arrowsraith, Ltd. (1 975)). The contents of these two documents are hereby incorporated...by characteristics a member must possess to satisfy the associated restriction. Examples of restrictions include **categories** such as proper name, location, country, date, unit of measurement, company name, baseball players, etc. Thus certain restrictions are best defined by lists of terms that fall into certain **categories** . Certain restrictions may be expressed as a morphological criterion, as a syntactic criterion, or as...the unit slow in the words slowly, slowest, etc. A morphological restriction may indicate a **category** such as a part of speech (e.g., noun, verb, preposition, adjective). Words falling within the **category** satisfy the restriction. For example, the three phrases Microsoft wins, Microsoft loses, and Microsoft settles...

...query Microsoft
[VEW since
wins, loses, and settles each fall within the part of speech **category** VERB and therefore satisfy the restriction that the partially unspecified term be a verb. A...

...unit from which sentences are constructed. Examples of restrictions expressed as syntactic criteria include phrasal **categories** such as noun phrase, verb phrase, or prepositional phrase. Fig. 11 shows an example of ...

...moon, etc. match a partially unspecified term with the restriction [PREPOSITIONAL PHRASE].

In addition to **categories** , morphological criteria, and syntactic criteria, a restriction can also be expressed as a computer program... module 44 may simply compare a term T that occurs within a document with predefined **category** lists 48 that include terms that satisfy the restriction to determine whether term T satisfies the restriction.

70
appearing within a plurality of documents, inforination indicating **category**
restrictions that the terms and contexts satisfy, and identifiers of the documents and contexts containing...

...is represented by one or more at least partially unspecified terms 1 1 reflecting a **category** restriction; (ii) identify preanalyzed contexts and
documents in the index that contain the one or...

7/5,K/6 (Item 6 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT

00848502 **Image available**
SYSTEM FOR FULFILLING AN INFORMATION NEED
SYSTEME REPONDANT A UN BESOIN D'INFORMATION
Patent Applicant/Assignee:
GLOBAL INFORMATION RESEARCH AND TECHNOLOGIES LLC, c/o Teragram Corp., 236 Huntington Avenue, Boston, MA 02115-4701, US, US (Residence), US (Nationality)
Inventor(s):
ROCHE Emmanuel, Gobal Information Research and Technologies LLC, c/o Teragram, 236 Huntington Avenue, Boston, MA 02215-4701, US,
SCHABES Yves, Global Information Research and Technologies LLC, c/o Teragram, 236 Huntington Avenue, Boston, MA 02115-4701, US,
Legal Representative:
GERBER Monica R (agent), Choate, Hall & Stewart, Exchange Place, 53 State Street, Boston, MA 02109, US,
Patent and Priority Information (Country, Number, Date):
Patent: WO 200182114 A2-A3 20011101 (WO 0182114)
Application: WO 2001US13150 20010424 (PCT/WO US0113150)
Priority Application: US 2000559223 20000426
Designated States:
(Protection type is "patent" unless otherwise stated - for applications prior to 2004)
AE AG AL AM AT AU AZ BA BB BG BR BY BZ CA CH CN CO CR CU CZ DE DK DM DZ
EE ES FI GB GD GE GH GM HR HU ID IL IN IS JP KE KG KP KR KZ LC LK LR LS
LT LU LV MA MD MG MK MN MW MX MZ NO NZ PL PT RO RU SD SE SG SI SK SL TJ
TM TR TT TZ UA UG UZ VN YU ZA ZW
(EP) AT BE CH CY DE DK ES FI FR GB GR IE IT LU MC NL PT SE TR
(OA) BF BJ CF CG CI CM GA GN GW ML MR NE SN TD TG
(AP) GH GM KE LS MW MZ SD SL SZ TZ UG ZW
(EA) AM AZ BY KG KZ MD RU TJ TM
Main International Patent Class: G06F-017/30
Publication Language: English
Filing Language: English
Fulltext Availability:
Detailed Description
Claims
Fulltext Word Count: 26836

English Abstract
The invention offers new approaches to fulfilling an information need, in particular to finding a result for a query based on a large body of information such as a collection of documents. The invention accepts a query containing an unspecified portion that expresses the information need. The invention locates matches for the query within a body of information and returns the matches or portions thereof in addition to or

instead of identifiers for documents in which the matches are found. The invention ranks the matches in order to provide the most relevant information. One preferred method of ranking considers the number of instances of a match among a plurality of documents. The invention further defines a new type of index that includes contexts in which terms occur and provides methods of searching such indices to fulfill an information need.

French Abstract

L'invention concerne de nouvelles approches de reponse a un besoin d'information permettant, en particulier, de trouver une reponse a une demande fondee sur une grande quantite d'informations, telles qu'une collection de documents. L'invention permet d'accepter une demande contenant une partie non specifiee exprimant un besoin information. Elle permet de localiser des correspondances entre la demande et un corps d'informations, et renvoie les correspondances ou des parties de celles-ci avec des identificateurs de documents dans lesquels des correspondances ont ete trouvees ou a la place de ceux-ci. Elle permet de classer les correspondances de facon a fournir les informations les plus pertinentes. Un procede de classement prefere, considere le nombre d'instances d'une correspondance dans plusieurs documents. Elle permet egalement de definir un nouveau type d'index comprenant des contextes dans lesquels des termes apparaissent, et fournit des procedes de recherche d'indices permettant de repondre a un besoin d'information.

Legal Status (Type, Date, Text)
Publication 20011101 A2 Without international search report and to be republished upon receipt of that report.
Examination 20020321 Request for preliminary examination prior to end of 19th month from priority date
Search Rpt 20030912 Late publication of international search report
Republication 20030912 A3 With international search report.
Republication 20030912 A3 Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

Fulltext Availability:
Detailed Description

Detailed Description

... partially unspecified, in which case (inverted exclamation mark)t includes a restriction that defines a **category** of ternis that are acceptable matches for the unspecified portion. The restriction may indicate that an unspecified portion of a query must meet a **category** criterion, a morpliological or syntactic criterion, or a criterion defined by a
7
computer program...26 and 27. In certain embodiments of the invention data stored in mernory 15 includes **category** lists 28, dictionaries 29, and index or indices 50 which are discussed further below.

Applications...can also be stored on peripheral storage 43. Data stored in memory 42 preferably includes **category** lists 48 and dictionaries 49. In preferred embodiments of the invention these are identical to **category** lists 28 and dictionaries 29 stored on query server(s) 4. In certain embodiments of the invention query servers 4 and indexing computer 6 access the same **category** lists and dictionaries. Index or indices 50 generated by indexing computer 6 are also preferably...

...indexer described in "The SMART Retrieval System.: Experiments in Automatic

1 8
Document Processing" by **Gerald Salton** (Prentice-Hall, Inc. (1971>> and "A Theory of Indexing" also by **Gerald Salton** (J. W. Arrowsmith, Ud. (1975>>. The contents of these two documents are hereby incorporated by...by characteristics a member must possess to satisfy the associated restriction. Examples of restrictions include **categories** such as proper name, location, country, date, unit of measurement, company name, etc. Thus certain restrictions are best defined by lists of terins that fall into certain **categories** . Certain restrictions may be expressed as a morpliological criterion, as a syntactic criterion, or as...

...the unit slow in the words slowly, slowest, etc. A morpliological restriction may indicate a **category** such as a part of speech (e.g., noun, verb, preposifion, adjective). Words falling within the **category** satisfy the restriction. For example, the three plirases Microsoft wins, Microsoft loses, and Mierosoft settles...

...query Microsoft
[VERBI since wins, loses, and settles each fall.

witIun the part of speech **category** VERB and therefore satisf`y the restriction that the partially unspecified tenn be a verb...

...unit from which sentences are constructed. Examples of restrictions expressed as syntactic criteria include plirasal **categories** such as noun plirase, verb plirase, or prepositional phrase. Figure 1 1 shows an example...

...moon, etc. match a partially unspecified tenn with. the restriction [PREPOSITIONAL PHRASE].

In addition to **categories** , morpliological criteria, and. syntactic criteria, a restriction can also be expressed as a computer program... satisfy the restriction to deterimne whether term T satisfies the restriction.

Depending upon the restriction, **category** lists 48 may be prepared by a human being.

(Note that such lists need not...

...49 containing infonnation regarding word fonn, part of speech, compound words and phrases, etc.

Preferably **category** lists 48 and dictionaries 49 as are needed by indexer module 44 are stored in...FSA for a context, information indicating which portions of the context satisfy the various restrictions (**categories** , morphological., syntactic, etc.) discussed above is incorporated. This task is perfonned by **category** recognition module 54, compound words and lexical phrases module 56, morphology module 58, and syntactic...

...has created FSA 53 for the terms in a context, FSA 53 is provided to **category** recognition module 54. **Category** recognition module 54 identifies terms that satisfy various restrietions that are defined as **categories** (e.g.,
COUNTRY, CITY, COMPANY NAME) and adds a new are labeled with the restriction for each term or group of terms that satisfies the restriction. **Category** recognition module 54 uses stored **category** lists 28 described above for this purpose.

FSA 55, generated by **category** recognition module 54, is provided to compound generated by morphology module 58, is then input...in the group tenninates. FSA 61, generated by syntactic phrase module 60, includes ares for
category restrictions, compound words and lexical phrases, morphological
restrictions, and syntactic restrictions satisfied by a terra...
...FSA, step 1340 preferably comprises the construction of the FSA by automaton generation module 52, **category** recognition module 54, compound words and lexical phrases module 56, morphology module 58, and syntactic...

...the FSA generated by automaton generation module 52 to include the affitional ares. In addition, **category** recognition module 54, compound words and lexical plirases' module 56, morphology module 58, and syntactic...

7/5,K/7 (Item 7 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT

00545199 **Image available**
SEARCH AND INDEX HOSTING SYSTEM
SYSTEME DE RECHERCHE ET D'HEBERGEMENT D'INDEX
Patent Applicant/Assignee:
GLOBAL INFORMATION RESEARCH AND TECHNOLOGIES LLC,
Inventor(s):
SCHABES Yves,
ROCHE Emmanuel,
BROWN Ryan,
Patent and Priority Information (Country, Number, Date):
Patent: WO 200008572 A1 20000217 (WO 0008572)
Application: WO 99US17359 19990802 (PCT/WO US9917359)
Priority Application: US 98130420 19980806
Designated States:
(Protection type is "patent" unless otherwise stated - for applications prior to 2004)
AE AL AM AT AU AZ BA BB BG BR BY CA CH CN CU CZ DE DK EE ES FI GB GD GE GH GM HR HU ID IL IN IS JP KE KG KP KR KZ LC LK LR LS LT LU LV MD MG MK MN MW MX NO NZ PL PT RO RU SD SE SG SI SK SL TJ TM TR TT UA UG UZ VN YU ZA ZW GH GM KE LS MW SD SL SZ UG ZW AM AZ BY KG KZ MD RU TJ TM AT BE CH CY DE DK ES FI FR GB GR IE IT LU MC NL PT SE BF BJ CF CG CI CM GA GN GW ML MR NE SN TD TG
Main International Patent Class: G06F-017/30
Publication Language: English
Fulltext Availability:
Detailed Description
Claims
Fulltext Word Count: 12008

English Abstract
The system initiates a search at a first network site for user-specified data in a remote database at a second network site and conducts the search at a third network site (e.g., at a host computer's site). To begin, the system receives, at the first network site, a provider identifier associated with the database from the second network site. Thereafter, the user-specified data is input at the first network site, following which the user-specified data and the provider identifier are output from the first network site to the third network site. The system

then searches for the user-specified data in a database at the third network site using the provider identifier. This database at the third network site includes data that corresponds to data stored in the remote database at the second network site.

French Abstract

Le systeme lance sur un premier site de reseau une recherche de donnees propres a un utilisateur dans une base de donnees eloignee d'un deuxieme site de reseau, puis effectue la recherche sur un troisieme site de reseau (par exemple celui d'un ordinateur hote). Au commencement, le systeme recoit au premier site de reseau l'identification d'un prestataire associee a une base de donnees du deuxieme site de reseau. Puis, les donnees propres a l'utilisateur sont introduites dans le premier site de reseau et transmises, avec l'identificateur de prestataire du premier site de reseau au troisieme site de reseau. Le systeme recherche alors les donnees propres a l'utilisateur dans la base de donnees du troisieme site de reseau a l'aide de l'identificateur de prestataire. Ladite base de donnees comprend des donnees correspondant aux donnees stockees dans la base de donnees distante du deuxieme site de reseau.

Fulltext Availability:
Detailed Description
Claims

Detailed Description
... engine and indexer
described in "The SMART Retrieval System: Experiments in Automatic Document Processing" by **Gerald Salton** (Prentice-Hall, Inc. (1971)) and
"A Theory of Indexing" also by **Gerald Salton** (J. W. Arrowsmith, Ltd.

(1975)). The contents of these two documents are hereby incorporated by ...

Claim
... AND EMPLOYMENT MATTERS AND DEVELOPIN(
INTERNATIONAL SEARCH REPORT
nternational Application No
PCTAS 99/17359
A. **CLASSIFICATION** OF SUBJECT MATTER
IPC 7 G06F17/30
According to International Patent **Classification** (IPC) or to both national **classification** and IPC
B. FIELDS SEARCHED
Minimum documentation searched (**classification** system followed by **classification** symbols)
I PC 7 G06F
Documentation searched other than minimum documentation to the extent that...

...of data base and, where practical, search terms used)
C. DOCUMENTS CONSIDERED TO BE RELEVANT
Category Citation of document, with indication, where appropriate, of the relevant passages Relevant to claim No...
...listed in the continuation of box C. Patent family members are listed in annex.
Special **categories** of cited documents :
'7" later document published after the international filing date or priority date...SEARCH REPORT
nternational Application No

PCTAS 99/17359
C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT
Category Citation of document, with indication,where appropriate, of the relevant passages Relevant to claim No...

7/5,K/8 (Item 8 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT

00457896 **Image available**
METHOD AND APPARATUS FOR PROCESSING FREE-FORMAT DATA
PROCEDE ET APPAREIL POUR TRAITER DES DONNEES NON STRUCTUREES
Patent Applicant/Assignee:
HETHERINGTON Greg,
Inventor(s):
HETHERINGTON Greg,
Patent and Priority Information (Country, Number, Date):
Patent: WO 9848360 A1 19981029
Application: WO 98AU288 19980422 (PCT/WO AU9800288)
Priority Application: AU 97439 19970422
Designated States:
(Protection type is "patent" unless otherwise stated - for applications prior to 2004)
AL AM AT AU AZ BA BB BG BR BY CA CH CN CU CZ DE DK EE ES FI GB GE GH GM GW HU ID IL IS JP KE KG KP KR KZ LC LK LR LS LT LU LV MD MG MK MN MW MX NO NZ PL PT RO RU SD SE SG SI SK SL TJ TM TR TT UA UG US UZ VN YU ZW GH GM KE LS MW SD SZ UG ZW AM AZ BY KG KZ MD RU TJ TM AT BE CH CY DE DK ES FI FR GB GR IE IT LU MC NL PT SE BF BJ CF CG CI CM GA GN ML MR NE SN TD TG
Main International Patent Class: G06F-017/30
International Patent Class: G06F-17:20
Publication Language: English
Fulltext Availability:
Detailed Description
Claims
Fulltext Word Count: 20141

English Abstract
A method and apparatus for processing free-format data (301) to produce a "text object" associated with the free-format data. The text object comprises a plurality of "component nodes" (302-312) containing attribute-type identifiers for elements of the free-format text and other data facilitating access to the text object to obtain information and/or change or add the free-format data. This arrangement obviates the need for the provision of separate database fields for each element of the information. Free-format data can therefore be processed in a similar manner to the way a human being processes free-format data. All elements can be accessed via the constructed text object.

French Abstract
L'invention concerne un procede et un appareil pour traiter des donnees non structurees (301) pour produire un "objet texte" associe a des donnees non structurees. Cet objet texte comprend plusieurs "noeuds composants" (302-312) contenant des identificateurs de type attributs pour des elements du texte non structure et d'autres donnees permettant d'acceder a l'objet texte pour obtenir des informations et/ou modifier ou ajouter les donnees non structurees. Cette configuration supprime la necessite de prevoir des chambres de bases de donnees distincts pour chaque element d'information. Les donnees non structurees peuvent ainsi etre traitees de la meme maniere qu'un etre humain traite des donnees non

structurees. Il est possible d'acceder a tous les elements via l'objet texte construit.

Fulltext Availability:
Detailed Description

Detailed Description
... performed internally by data management systems.

Generally, data management systems may be divided into two **categories** : 1) Database management systems; and 2) Text search and retrieval systems.

The first type of...
...representation of this data
which is read by humans or used by another system. This **category** of data management system includes: hierarchical, network, relational, object-oriented database management systems and knowledge...

...information about many entities).

Within each record the various "attributes" of the entity are usually **classified** into "fields".

Within object-oriented database management systems and knowledge based management systems these basic...Systems".

They use context--free rewrite rules with non terminal semantic constituents. The constituents are **categories** or metasymbols such as attribute, object, present (as in display or print), and ship, rather...small change would require the whole text string to regenerated.

The text search and retrieval **category** of data management system does not import the data but builds searchable indices which point to the original data. This **category** includes: document storage & retrieval systems; and Internet search engines.

These types of systems have very...

...include.

Cross checking and validating the data
Integrating the data with database systems
Sorting and **classifying** the text data
From these limitations, we can see that this **category** of data management system is suited to unstructured data which does not need to be...

...look at words in context. Some others identify words that are nouns but do not **classify** the type of ...changed within context.

For more information regarding this area, refer to the works published by **Gerald Salton** .

Note that the term "text object" as used in the following description should not be...a predetermined hierarchy. Each component node may comprise.

0 an attribute type identifier (for the **classification** of an attribute of the free-format data which is associated with that component node...the embodiment can "imply" that it is a <StreetName>. Therefore, "12 Pitt Street" can be **classified** as a <Street> from the relative positioning of the tokens.

Domain Object
The main function...All of these priorities are specified in the input grammar file 1603 (figure 16).

0 **Classifying** the component nodes of the syntax tree as either visible or invisible. Low level "regular expression" terms such as <word> are **classified** as invisible.

0 Assigning match weightings to all component nodes. These values are specified in...method in two major ways. 1) All constituent parts of the free-format text are **classified** and used to reference the index. (i.e. not just the nouns). 2) There are...on "key word searching" techniques, refer to the numerous books and journal articles published by **Gerald Salton** .

"Key word search" techniques-applicable to this invention include.

0 Storing very common terms in...

7/5,K/9 (Item 9 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT

00311300 **Image available**
SYSTEM FOR DIRECTING RELEVANCE-RANKED DATA OBJECTS TO COMPUTER USERS
SYSTEME PERMETTANT DE DIRIGER DES DONNEES CLASSEES SELON LEUR PERTINENCE A DES UTILISATEURS D'ORDINATEURS
Patent Applicant/Assignee:
APPLE COMPUTER INC,
ROSE Daniel E,
BORNSTEIN Jeremy J,
TIENE Kevin,
PONCELEON Dulce B,
Inventor(s):
ROSE Daniel E,
BORNSTEIN Jeremy J,
TIENE Kevin,
PONCELEON Dulce B,
Patent and Priority Information (Country, Number, Date):
Patent: WO 9529452 A1 19951102
Application: WO 95US5072 19950424 (PCT/WO US9505072)
Priority Application: US 94231656 19940425
Designated States:

(Protection type is "patent" unless otherwise stated - for applications prior to 2004)
AM AT AU BB BG BR BY CA CH CN CZ DE DK EE ES FI GB GE HU IS JP KE KG KP KR KZ LK LR LT LU LV MD MG MN MW MX NO NZ PL PT RO RU SD SE SG SI SK TJ TM TT UA UG US UZ VN KE MW SD SZ UG AT BE CH DE DK ES FR GB GR IE IT LU MC NL PT SE BF BJ CF CG CI CM GA GN ML MR NE SN TD TG

Main International Patent Class: G06F-017/30

Publication Language: English

Fulltext Availability:
Detailed Description
Claims

Fulltext Word Count: 6614

English Abstract

An information access system stores items of information in an unstructured global database. When a user requests access to the system, the system delivers to that user an identification of only those items of information which are believed to be relevant to the user's interest. The determination as to the items of information that are relevant to a user is carried out by ranking each available item in accordance with any one or more techniques. In one approach, the content of each document is matched with an adaptive profile of a user's interest. In another approach, a feedback mechanism is provided to allow users to indicate their degree of interest in each item of information. These indications are used to determine whether other users, who have similar or dissimilar interests, will find a particular item to be relevant.

French Abstract

Systeme d'acces a des informations qui met en memoire des elements d'information dans une base de donnees globale non structuree. Lorsqu'un utilisateur demande l'acces au systeme, ledit systeme fournit a cet utilisateur une identification des seuls elements d'information consideres comme pertinents pour les interets de l'utilisateur. La determination de la pertinence des elements d'information pour un utilisateur est effectuee par classement de chaque element disponible selon une ou plusieurs techniques. Dans une approche proposee, le contenu de chaque document est compare a un profil adaptatif des interets d'un utilisateur. Dans une autre approche, un mecanisme de reaction permet aux utilisateurs d'indiquer leur degre d'interet pour chaque element d'information. Ces indications sont utilisees pour determiner si d'autres utilisateurs qui ont des interets similaires ou differents trouveront pertinent un element particulier d'information.

Fulltext Availability:
Detailed Description

Detailed Description

... any user in a group to which the system pertains. The information need not be **classified** by topic or addressed to specific mailboxes or other user designations. In other words, each...described as being "unstructured" to denote
the fact that the messages stored therein are not **classified** under different topic **categories** or otherwise arranged in a structured manner that requires a user to designate a navigation...vector models for indexing text, reference is made to
Introduction To Modern Information Retrieval by **Gerald** **Salton** and Michael J.

McGill (McGraw-Hill 1983), which is incorporated herein by reference.

Each user...items of information be addressed to specific users, or

requiring the users to specifically select **categories** of interest, all available items of information are ranked in accordance with a predicted degree...particular message to be of interest, as in electronic mail systems, or into which topical **category** it should be placed. Similarly, recipients do not have to determine where to look to...are stored in a single global database 22. If desired, additional databases directed to specific **categories** of information can be included. For example, a database of movie descriptions can be provided...

7/5,K/10 (Item 10 from file: 349)

00311299 **Image available**

SYSTEM FOR RANKING THE RELEVANCE OF INFORMATION OBJECTS ACCESSED BY COMPUTER USERS

SYSTEME PERMETTANT DE CLASSER PAR ORDRE D'IMPORTANCE LA PERTINENCE D'OBJETS D'INFORMATION AUXQUELS ONT ACCES DES UTILISATEURS D'ORDINATEURS

Patent Applicant/Assignee:
APPLE COMPUTER INC,
ROSE Daniel E,
BORNSTEIN Jeremy J,
TIENE Kevin,
PONCELEON Dulce B,
Inventor(s):
ROSE Daniel E,
BORNSTEIN Jeremy J,
TIENE Kevin,
PONCELEON Dulce B,

Patent and Priority Information (Country, Number, Date):
Patent: WO 9529451 A1 19951102
Application: WO 95US5070 19950424 (PCT/WO US9505070)
Priority Application: US 94231655 19940425
Designated States:
(Protection type is "patent" unless otherwise stated - for applications prior to 2004)
AM AT AU BB BG BR BY CA CH CN CZ DE DK EE ES FI GB GE HU IS JP KE KG KP KR KZ LK LR LT LU LV MD MG MN MW MX NO NZ PL PT RO RU SD SE SG SI SK TJ TM TT UA UG US UZ VN KE MW SD SZ UG AT BE CH DE DK ES FR GB GR IE IT LU MC NL PT SE BF BJ CF CG CI CM GA GN ML MR NE SN TD TG
Main International Patent Class: G06F-017/30
Publication Language: English
Fulltext Availability:
Detailed Description
Claims
Fulltext Word Count: 5826

English Abstract
Information presented to a user via an information access system is ranked according to a prediction of the likely degree of relevance to the user's interests. A profile of interests is stored for each user having access to the system. Items of information to be presented to a user are ranked according to their likely degree of relevance to that user and displayed in order of ranking. The prediction of relevance is carried out by combining data pertaining to the content of each item of information with other data regarding correlations of interests between users. A value indicative of the content of a document can be added to another value which defines user correlation, to produce a ranking score for a document. Alternatively, multiple regression analysis or evolutionary programming can be carried out with respect to various factors pertaining

to document content and user correlation, to generate a prediction of relevance. The user correlation data is obtained from feedback information provided by users when they retrieve items of information. Preferably, the user provides an indication of interest in each document which he or she retrieves from the system.

French Abstract
Des informations presentees a un utilisateur par l'intermediaire d'un systeme d'acces aux informations sont classees selon une prevision de leur degre vraisemblable de pertinence au vu des interets de l'utilisateur. Un profil d'interets est mis en memoire pour chaque utilisateur ayant acces au systeme. Des elements d'information a presenter a un utilisateur sont classes selon leur degre vraisemblable de pertinence pour ledit utilisateur et affiches selon l'ordre de classement. La prevision de pertinence est effectuee par combinaison de donnees relatives au contenu de chaque element d'information avec d'autres donnes relatives a des correlations d'interets entre les utilisateurs. Une valeur indicatrice du contenu d'un document peut etre ajoutee a une autre valeur qui definit une correlation entre utilisateurs, pour produire un chiffre de classement pour ledit document. Dans une variante, une analyse de regression multiple ou une programmation evolutive peuvent etre effectuees sur la base de divers facteurs relatifs au contenu du document et a la correlation entre utilisateurs, pour produire une prevision de la pertinence. Les donnees de correlation entre utilisateurs sont obtenues a partir de reactions communiquees par les utilisateurs lorsqu'ils recuperent des elements d'information. De preference, l'utilisateur fournit une indication de son interet pour chaque document qu'il recupere dans le systeme.

Fulltext Availability:
Detailed Description
Claims

Detailed Description
... undesirable mail.

Similarly, in bulletin board systems, the number of documents in a particular topical **category** at any given time can be quite significant. The user must try to identify documents...
...be missed if the user cannot take the time to view all documents in the **category** .

Along similar lines, in a text retrieval system, a broadly framed query can result in...e.g., messages in an electronic mail network or documents within a particular bulletin board **category** , are ranked according to their likely degree of relevance and displayed with an indication of ... can be stored in a single database 22. If desired, multiple databases directed to specific **categories** of information can be provided. For example, a separately accessible database of movie descriptions can... into account the type of movie, such as action or drama, the actors, its viewer **category** rating, and the like.

The example of Figure 5A illustrates a two-dimensional vector for... vector models for indexing text, reference is made to Introduction To Modem Information Retrieval by **Gerald Salton** and Michael J.

McGill (McGraw-Hill 1983), which is incorporated herein by reference.

Each user...

? t17/5,k/all

17/5,K/1 (Item 1 from file: 348)
DIALOG(R)File 348:EUROPEAN PATENTS

00883842
Hypertext document retrieving apparatus for retrieving hypertext documents relating to each other
Hypertext-Dokumentwiederauffindungssystem zum Wiederauffinden zusammengehoriger Hypertextdokumente
Systeme de recouvrement de documents hypertextes pour retrouver des documents hypertexte relies
PATENT ASSIGNEE:
MATSUSHITA ELECTRIC INDUSTRIAL CO., LTD., (216887), 1006, Oaza Kadoma, Kadoma-shi, Osaka-fu, (JP), (Proprietor designated states: all)
INVENTOR:
Ishikawa, Masato, 2-37-6, Horinouchi, Suginami-ku, Tokyo, (JP)
Sato, Mitsuhiro, 2-27-1-A-201, Hashido, Seya-ku, Yokohama, (JP)
Hoshida, Masaki, 6-10-9-101, Togoshi, Shinagawa-ku, Tokyo, (JP)
Noguchi, Yoshihiro, 1-11-13-203, Fukuei, Ichikawa-shi, Chiba-ken, (JP)
Yasukawa, Hideki, 3-5-8-101, Kichijoji, Kita-machi, Musashino-shi, Tokyo, (JP)
LEGAL REPRESENTATIVE:
Schmidt, Christian, Dipl.-Phys. et al (76643), Manitz, Finsterwald & Partner GbR Postfach 31 02 20, 80102 Munchen, (DE)
PATENT (CC, No, Kind, Date): EP 809197 A2 971126 (Basic)
EP 809197 A3 010214
EP 809197 B1 040204
APPLICATION (CC, No, Date): EP 97107823 970513;
PRIORITY (CC, No, Date): JP 96149783 960522
DESIGNATED STATES: DE; FR; GB
INTERNATIONAL PATENT CLASS: G06F-017/30
CITED PATENTS (EP B): EP 679999 A; WO 95/00896 A
CITED REFERENCES (EP B):
FRISSE M E: "SEARCHING FOR INFORMATION IN A HYPERTEXT MEDICAL HANDBOOK" COMMUNICATIONS OF THE ASSOCIATION FOR COMPUTING MACHINERY,US,ASSOCIATION FOR COMPUTING MACHINERY. NEW YORK, vol. 31, no. 7, page 880-886 XP000051078 ISSN: 0001-0782
SALTON G ET AL: "AUTOMATIC STRUCTURING AND RETRIEVAL OF LARGE TEXT FILES" COMMUNICATIONS OF THE ASSOCIATION FOR COMPUTING MACHINERY,US,ASSOCIATION FOR COMPUTING MACHINERY. NEW YORK, vol. 37, no. 2, page 97-108 XP000425939 ISSN: 0001-0782
DUNLOP M D ET AL: "HYPERMEDIA AND FREE TEXT RETRIEVAL" INFORMATION PROCESSING & MANAGEMENT (INCORPORATING INFORMATION TECHNOLOGY),GB,PERGAMON PRESS INC. OXFORD, vol. 29, no. 3, page 287-298 XP002043306 ISSN: 0306-4573
WEISS R ET AL: "HYPURSUIT: A HIERARCHICAL NETWORK SEARCH ENGINE THAT EXPLOITS CONTENT-LINK HYPERTEXT CLUSTERING" ACM CONFERENCE ON HYPERTEXT,US,NEW YORK, ACM, vol. CONF. 7, page 180-193 XP000724328 ISBN: 0-89791-778-2;

ABSTRACT EP 809197 A2
A hypertext document and anchor sentences of parent documents for the hypertext document are registered with an hypertext document identifier as document information for each of hypertext documents having reference relationships with each other. A user can refer to one hypertext document according to an anchor sentence of another hypertext document functioning as a parent document. Also, occurrence positions of one word in hypertext documents and parent documents are registered as word information for

each of words. When a keyword is input, a plurality of particular hypertext documents and particular parent documents in which the keyword appears are specified according to the word information, one particular hypertext document and corresponding particular parent documents are unified to a unified hypertext document for each particular hypertext document, an occurrence **frequency** of the keyword in each unified hypertext document is calculated according to the document information, importance degrees of the unified hypertext documents are calculated as those of the particular hypertext documents according to the occurrence **frequencies** , and ranking of the particular hypertext documents are determined according to those importance degrees. Because the occurrence **frequency** is calculated by considering the parent documents, the particular hypertext documents can be appropriately ranked.

ABSTRACT WORD COUNT: 198

NOTE:

Figure number on first page: 3

LEGAL STATUS (Type, Pub Date, Kind, Text):

Search Report:	010214	A3	Separate publication of the search report
Application:	971126	A2	Published application (A1with Search Report ;A2without Search Report)
Oppn None:	050126	B1	No opposition filed: 20041105
Grant:	040204	B1	Granted patent
Examination:	971126	A2	Date of filing of request for examination: 970513

LANGUAGE (Publication,Procedural,Application): English; English; English

FULLTEXT AVAILABILITY:

Available Text	Language	Update	Word Count
CLAIMS A	(English)	199711W3	2086
CLAIMS B	(English)	200406	2113
CLAIMS B	(German)	200406	1818
CLAIMS B	(French)	200406	2376
SPEC A	(English)	199711W3	14993
SPEC B	(English)	200406	14994
Total word count - document A			17082
Total word count - document B			21301
Total word count - documents A + B			38383

...ABSTRACT documents are unified to a unified hypertext document for each particular hypertext document, an occurrence **frequency** of the keyword in each unified hypertext document is calculated according to the document information...

...hypertext documents are calculated as those of the particular hypertext documents according to the occurrence **frequencies** , and ranking of the particular hypertext documents are determined according to those importance degrees. Because the occurrence **frequency** is calculated by considering the parent documents, the particular hypertext documents can be appropriately ranked.

...SPECIFICATION retrieval index developing unit 202 appears in each of the documents. That is, an occurrence **frequency** of each word in one document is calculated for each of the documents stored in...

...as a correction factor for the word for each of the words, a normalized occurrence **frequency** (called a TF value) of each word is calculated for each of the documents, an...

...calculated for each of the words by multiplying the deviation degree and the normalized occurrence **frequency** together, and a retrieval index is

developed in the retrieval index developing unit 202. In...

...the number of documents stored in the document managing unit 201. Also, the normalized occurrence **frequency** TF (=Fo/Nwd) obtained by dividing an occurrence **frequency** Fo of the remarked word in a remarked document by the number Nwd of words...

...estimated value TF*IDF is calculated by multiplying the deviation degree and the normalized occurrence **frequency** together.

The detail of the estimated value TF*IDF and a conventional document retrieving apparatus in which the estimated value TF*IDF is used are disclosed in a literature " **Salton** , **Gerard** : Introduction to modern Information Retrieval, McGraw-Hill computer science series, 1983).

2.2. PROBLEMS TO...

...unified to a unified hypertext document for each of the particular hypertext documents, an occurrence **frequency** of the keyword in one unified hypertext document is calculated for each unified hypertext document...

...plurality of importance degrees of the unified hypertext documents are determined according to the occurrence **frequencies** in the unified hypertext documents, one importance degree of one unified hypertext document is set...each of the particular hypertext documents obtained in the retrieving unit 3, calculating an occurrence **frequency** TF of one particular word in one unified particular hypertext document for each particular word and each unified particular hypertext document, calculating an inverse document **frequency** IDF defined as an inverse value of the number of particular hypertext documents, in which...

...particular word appears, for each particular word, calculating a product TF*IDF of one occurrence **frequency** TF and one inverse document **frequency** IDF, summing a plurality of products for all particular words to produce a summed product...

...the number of occurrence documents in which a word appears (generally called an inverse document **frequency** IDF) and the occurrence **frequency** of the word in each of the occurrence documents (generally called a text **frequency** TF) be calculated in advance in the retrieval index preparing unit 6 and written in...

...by the occurrence document identifiers transmitted from the retrieving unit 3. Thereafter, an inverse document **frequency** IDF defined as an inverse value of the number of unified particular hypertext documents in which one particular word agreeing with one keyword appears and the occurrence **frequency** TF of one particular word in each of the unified particular hypertext documents are calculated...

...plurality of sets of the occurrence document identifiers and the positional information. The inverse document **frequency** IDF denotes a correction factor for each particular word.

Thereafter, in cases where one keyword is only input, an estimated value obtained by multiplying the inverse document **frequency** IDF for one particular word and the occurrence **frequency** TF together is calculated as an importance degree for each of the unified particular hypertext...

...input by the user is two or more, a product TF*IDF of one occurrence **frequency** TF and one inverse document **frequency** IDF is calculated for

de classement des documents, une **frequence** d'occurrence de chaque mot apparente dans les documents hypertexte particuliers de rang superieur est...

...cle, une pluralite des degres d'importance des mots apparentes est calculee a partir des **frequences** d'occurrence des mots apparentes par le moyen de determination de classement de documents, un...

...hypertexte particuliers de rang superieur par le moyen de determination de classement de documents, une **frequence** d'occurrence de chaque mot apparente dans les documents hypertexte particuliers de rang superieur et...
...cle, une pluralite des degres d'importance des mots apparentes est calculee a partir des **frequences** d'occurrence des mots apparentes par le moyen de determination de classement de documents, un...

...pluralite de mots-cle est recue par le moyen de reception de mot-cle, une **frequence** d'occurrence TF d'un mot-cle dans un document hypertexte unifie est calculee par...

...determination de classement de document pour chaque mot-cle et chaque document hypertexte unifie, une **frequence** de document inverse IDF definie comme une valeur inverse du nombre de documents hypertexte particuliers...

...determination de classement de documents pour chaque mot-cle, un produit TF*IDF d'une **frequence** d'occurrence TF et d'une **frequence** de document inverse IDF est calculee par le moyen de determination de classement de documents...

17/5,K/2 (Item 2 from file: 348)
DIALOG(R)File 348:EUROPEAN PATENTS

00804220
Method and apparatus for training a text classifier
Verfahren und Gerat, um einen Textklassifizierer zu trainieren
Procede et dispositif d'apprentissage pour un classificateur **de textes**
PATENT ASSIGNEE:
AT&T IPM Corp., (1907680), 2333 Ponce de Leon Boulevard, Coral Gables, Florida 33134, (US), (applicant designated states: DE;FR;GB)
INVENTOR:
Lewis, David Dolan, 851 Springfield Avenue, Apt. 10G, Summit, New Jersey 08901, (US)
LEGAL REPRESENTATIVE:
Watts, Christopher Malcolm Kelway, Dr. et al (37391), Lucent Technologies (UK) Ltd, 5 Mornington Road, Woodford Green Essex, IG8 0TU, (GB)
PATENT (CC, No, Kind, Date): EP 747846 A2 961211 (Basic)
APPLICATION (CC, No, Date): EP 96303840 960529;
PRIORITY (CC, No, Date): US 484436 950607
DESIGNATED STATES: DE; FR; GB
INTERNATIONAL PATENT CLASS: G06F-017/30;

ABSTRACT EP 747846 A2
A method and apparatus for training a text **classifier** is disclosed. A supervised learning system and an annotation system are operated cooperatively to produce a **classification** vector which can be used to **classify** documents with respect to a defined class. The annotation system automatically annotates documents with a degree of relevance

annotation to produce machine annotated data. The degree of relevance annotation represents the degree to which the document belongs to the defined class. This machine annotated data is used as input to the supervised learning system. In addition to the machine annotated data, the supervised learning system can also receive manually annotated data and/or a user request. The machine annotated data, along with the manually annotated data and/or the user request, are used by the supervised learning system to produce a **classification** vector. In one embodiment, the supervised learning system comprises a relevance feedback mechanism. The relevance feedback mechanism is operated cooperatively with the annotation system for multiple iterations until a **classification** vector of acceptable accuracy is produced. The **classification** vector produced by the invention is the result of a combination of supervised and unsupervised learning. (see image in original document)

ABSTRACT WORD COUNT: 213

LEGAL STATUS (Type, Pub Date, Kind, Text):

Application:	961211	A2	Published application (A1with Search Report ;A2without Search Report)
Withdrawal:	971203	A2	Date on which the European patent application was withdrawn: 971006

LANGUAGE (Publication,Procedural,Application): English; English; English

FULLTEXT AVAILABILITY:

Available Text	Language	Update	Word Count
CLAIMS A	(English)	EPAB96	679
SPEC A	(English)	EPAB96	7038
Total word count - document A			7717
Total word count - document B			0
Total word count - documents A + B			7717

Method and apparatus for training a text classifier
Procede et dispositif d'apprentissage pour un classificateur **de textes**

...ABSTRACT A2

A method and apparatus for training a text **classifier** is disclosed. A supervised learning system and an annotation system are operated cooperatively to produce a **classification** vector which can be used to **classify** documents with respect to a defined class. The annotation system automatically annotates documents with a...

...and/or the user request, are used by the supervised learning system to produce a **classification** vector. In one embodiment, the supervised learning system comprises a relevance feedback mechanism. The relevance feedback mechanism is operated cooperatively with the annotation system for multiple iterations until a **classification** vector of acceptable accuracy is produced. The **classification** vector produced by the invention is the result of a combination of supervised and unsupervised ...

...SPECIFICATION A2

Field of the Invention

The present invention relates generally to computerized text **classification** . More particularly, the present invention relates to the combined supervised and unsupervised learning of text...

...ones which are not useful.

An important technique in on-line text processing is text **classification** , which is the sorting of documents into meaningful groups. A variety of text **classification** systems are currently in use.

in the **classifier** vector c. The c subscript identifies the weight term as a weight in the **classification** vector.

2.3 Document **Classification**

The **classification** vector c is used to rank the documents in a collection as follows. The **classification** vector c is applied to a document to calculate a retrieval status value (RSV) for...

...is computed according to the following equation: (Formula omitted) Thus, each weight term in the **classifier** vector W(sub(ck)) is multiplied by the corresponding weight term in the document vector...

...these multiplied weights gives a retrieval status value RSV, which represents the rank of the **classified** document. The higher the RSV, the more likely that the document falls within the class of documents represented by the **classification** vector c.

3. Invention Overview

A block diagram illustrating a system for training a **classifier** in accordance with the present invention is shown in Fig. 2. Fig. 2 shows a for training a **classifier** includes a supervised learning system 210, an automatic annotation system 220, and a document database...

...class of documents of interest to a user, and which are used to produce a **classification** vector c. This **classification** vector c can be used to **classify** documents in the database 230 with respect to a class of interest. The remainder of...

...relevant to that user, although the present invention is not limited to such classes.

The **classification** vector c produced by the supervised learning system 210 is input to the annotation system 220. The annotation system 220 **classifies** the documents in the database 230 using the **classification** vector c and automatically annotates the documents to produce machine annotated data. The machine annotated...

...to the supervised learning system 210 during subsequent iterations in order to produce a new **classification** vector c based upon both 1) the machine annotated data, and 2) the manually annotated data and/or user request. This procedure continues until a **classification** vector c of acceptable accuracy is produced. Thus, the supervised learning system 210 is capable...304. (alpha) controls how much weight the initial request T has during formation of the **classification** vector c as discussed in further detail in section 4.2.1 below. (beta) and (gamma) control how much weight to give relevant and non-relevant documents, respectively, during **classification** vector formation. These relevant and non-relevant documents may be documents that were manually annotated...

...gamma) =4, based on the discussion of the setting of these parameters in Chris Buckley, **Gerard Salton** , and James Allan, "The Effect Of Adding Relevance Information In A Relevance Feedback Environment", in...

...are relevant. These automatic initial P(sub(i)) annotations will change during processing.

4.2 **Classification** Vector Formation

The algorithm for producing a **classification** vector c in one embodiment is based on the Rocchio algorithm for relevance feedback, which...

...11 of Information Retrieval: Data Structures and Algorithms, Prentice-Hall Publishing, 1992. The algorithm for **classification** vector formation is as follows.

4.2.1 Construction of **Classification** Vector by Relevance Feedback Module

The **classification** vector c is constructed by the relevance feedback module 310 in step 430. The relevance...

...the iteration probability annotator 322 (described in further detail below) during subsequent iterations.

The constructed **classification** vector c is a vector (Formula omitted) where (see image in original document) according to...

...of the above equation are the same as the elements of the Rocchio formula for **classification** vector construction in relevance feedback. The first element (alpha)w(sub(rk)) increases the weight of the kth indexing term in the **classification** vector c in proportion to the weight the term has in the query vector. The...

...the term in the query has on the final weight of the term in the **classification** vector c.

The second element (see image in original document) increases the weight of the kth indexing term in the **classification** vector c in proportion to the average weight of the term in the documents which...

...in the manually annotated relevant documents has on the weight of the term in the **classification** vector.

The third element (see image in original document) decreases the weight of the kth indexing term in the **classification** vector c in proportion to the average weight of the term in the documents which...

...in the manually annotated nonrelevant documents has on the weight of the term in the **classification** vector

The last two elements (Formula omitted) modify the Rocchio formula by taking into account...

...fourth term in the equation increases the weight of the kth indexing term in the **classification** vector c according to the proportional average weight of the term in the machine annotated...

...fifth term in the equation decreases the weight of the kth indexing term in the **classification** vector c according to the proportional average weight of the term in the machine annotated...

...formula using similar methods.

Thus, the output of the relevance feedback module 310 is a **classification** vector c, which is formed by the above equation.

4.3 Operation of the Annotation...

...document.

4.3.1 Operation of Search Engine to Produce Document Retrieval Status Values

The **classification** vector c = <w(sub(c))(sub(1))...w(sub(ck))...w(sub(cd))> produced by...

...from the document database 230, are provided to the search engine 324, which performs the **classification** function in step 435. The **classification** vector c is applied to all of the documents, both

vector.
4. The method of claim 1 wherein said defined class is defined by a...

...said defined class.
8. A method for operation of a computer system for producing a **classification** vector for use in **classifying** documents with respect to a defined class, said method comprising the steps of:
calculating a...

...the relevance feedback function using a manually annotated document.
11. An apparatus for training a **classifier** to classic documents with respect to a defined class, said apparatus comprising:
an annotation system...

...one document belongs to said defined class; and
a supervised learning system for training the **classifier** using said at least one automatically annotated document.
12. The system of claim 11 wherein...

...relevance feedback module.
14. The system of claim 11 wherein said annotation system further comprises:
classification means for calculating a retrieval status value for said at least one document; and
means responsive to said **classification** mess for calculating said degree of relevance for said at least one document.
15. An apparatus for training a **classifier** to **classify** documents with respect to a defined class, said apparatus comprising:
means for calculating a degree...

...said at least one document belongs to said defined class; and
means for training the **classifier** using said calculated degree of relevance.
16. The apparatus of claim 15 further comprising:
annotation...

...to produce at least one automatically annotated document; and
wherein said means for training the **classifier** comprises a relevance feedback mechanism responsive to said annotation means for producing a **classification** vector.
17. The apparatus of claim 15 wherein said degree of relevance is proportional to...

17/5,K/3 (Item 3 from file: 348)
DIALOG(R)File 348:EUROPEAN PATENTS

00705119
ASSOCIATIVE TEXT SEARCH AND RETRIEVAL SYSTEM
ASSOZIATIVES TEXTSUCH- UND WIEDERAUFFINDUNGSSYSTEM
SYSTEME ASSOCIATIF DE RECHERCHE ET DE RECUPERATION DE TEXTE
PATENT ASSIGNEE:
LEXIS-NEXIS, A DIVISION OF REED ELSEVIER INC., (2092360), 9443 Springboro Pike, P.O. Box 933, Dayton, OH 45401, (US), (Proprietor designated states: all)
INVENTOR:
HOLT, John, 4405 Lac Lamen, Centerville, OH 45458, (US)
MILLER, David, James, 2689 Woodbluff Lane, Spring Valley, OH 45370, (US)
LU, Allan, X., 320 Brookside Drive, Springboro, OH 45066, (US)

DALEY, Ray, 3464 Cottage Point Way, Dayton, OH 45449, (US)
DOAN, Minh, 4141 Bronze Leaf Ct., Dayton, OH 45424, (US)
GRAHAM, Richard, G., 222 Kenderton Trail, Beavercreek, OH 45430, (US)
LEININGER, Catherine, 40 Fernwood Drive, Dayton, OH 45440, (US)
McBEATH, Darin, W., 10214 Forestedge Lane, Miamisburg, OH 45342, (US)
PEASE, Thomas, 750 Anthony Lane, Mason, OH 45040, (US)
SEVER, Stephen, M., 2724 Oak Park Avenue, Kettering, OH 45419, (US)
WADDELL, Dale, 2648 Greenstone Court, Dayton, OH 45430, (US)
WECKESSER, Franz, 67 Bizzell Avenue, Dayton, OH 45459, (US)

LEGAL REPRESENTATIVE:
Ede, Eric et al (61984), Fitzpatricks, 4 West Regent Street, Glasgow G2 1RS, (GB)

PATENT (CC, No, Kind, Date): EP 730765 A1 960911 (Basic)
EP 730765 A1 970917
EP 730765 B1 030917
WO 95014973 950601

APPLICATION (CC, No, Date): EP 95902589 941122; WO 94US13272 941122

PRIORITY (CC, No, Date): US 155304 931122

DESIGNATED STATES: AT; BE; CH; DE; DK; ES; FR; GB; GR; IE; IT; LI; LU; MC; NL; PT; SE

INTERNATIONAL PATENT CLASS: G06F-017/30

CITED PATENTS (EP B): WO 92/04681 A; US 4270182 A; US 4787035 A; US 5323316 A

CITED REFERENCES (EP B):
INFORMATION PROCESSING & MANAGEMENT, SEPT.-OCT. 1993, UK, vol. 29, no. 5, ISSN 0306-4573, pages 647-669, XP002035616 WONG W Y P ET AL: "Implementations of partial document ranking using inverted files"
U.S. DEPARTMENT OF COMMERCE, "Text Search and Retrieval Training Manual for the Automated patent System (APS)", 21 October 1992, pages (1-5), (3-2) to (3-3), (3-8) to (3-9), (3-15), (7-35) to (7-37), (7-1) to (7-7).
U.S. DEPARTMENT OF COMMERCE, "Text Search and Retrieval Reference Manual for the Automated Patent System (APS)", 21 October 1992, pages 28, 29, 39-43.
DOCTOR OF PHILOSOPHY, "Inference Networks for Document Retrieval", TURTLE, February 1991, page 65.;

NOTE:
No A-document published by EPO

LEGAL STATUS (Type, Pub Date, Kind, Text):

Examination:	010425 A1	Date of dispatch of the first examination report: 20010313
Application:	950830 A	International application (Art. 158(1))
Oppn None:	040908 B1	No opposition filed: 20040618
Grant:	030917 B1	Granted patent
Lapse:	040728 B1	Date of lapse of European Patent in a contracting state (Country, date): DE 20031218,
Application:	960911 A1	Published application (A1with Search Report ;A2without Search Report)
Examination:	960911 A1	Date of filing of request for examination: 960528
Change:	970625 A1	Representative (change)
Search Report:	970917 A1	Drawing up of a supplementary European search report: 970730

LANGUAGE (Publication,Procedural,Application): English; English; English

FULLTEXT AVAILABILITY:

Available Text	Language	Update	Word Count
CLAIMS B	(English)	200338	3091
CLAIMS B	(German)	200338	2760
CLAIMS B	(French)	200338	3578
SPEC B	(English)	200338	7206

Total word count - document A 0
Total word count - document B 16635
Total word count - documents A + B 16635

...SPECIFICATION number of searches.
Associative retrieval, a technique for information retrieval developed in the 1960s by **Gerard Salton** , addresses some of the shortcomings of Boolean searching. Automatic Text Processing, (published by Addison Wesley, New York, New York 1988, and written by **Gerard Salton**) provides a description of associative retrieval searching. The basic formula used in associative retrieval involves...

...that occur within the document. The two basic weighting factors are known as the term **frequency** - tf -and the inverse document **frequency** - idf.
The term **frequency** is defined as the number of times the term occurs within a given document. Hence, the term **frequency** must be calculated for each document within the collection.
The inverse document **frequency** is defined as the inverse of the number of documents in the entire collection which...

...each document using a ranking formula that varies according to the square of the term **frequency** of each of the search terms in the document. The ranking formula can also vary according to the inverse document **frequency** of each search term. The formula can also use a maximum term **frequency** to estimate the size of a document and the maximum document **frequency** to estimate the number of documents in a collection of documents, thus reducing the amount...

...distinguishing between noise words, which are not provided in an index for the documents, and **frequently** used terms, which are provided in the index but which are not used in the...

...indicates the importance of each term, which varies according to the inverse of the document **frequency** of each term. The documents can be sorted according to rank or according to a...two terms 173 "FIRST" and "CASE". The asterisk indicates that the terms which follow are " **frequently** used terms". Any terms that are deemed **frequently** used terms are eliminated from further processing for the search because their value in locating...

...search illustrated by the screen 170, the two terms 173 "FIRST" and "CASE" were deemed **frequently** used terms and hence were not searched. The ancillary data 76, shown in connection with FIG. 3, contains a table of **frequently** used terms for each physical document collection. The determination as to which terms go into...

...known to one of ordinary skill in the art, including, but not limited to, the **frequency** of occurrence of a term in the physical collection and the relationship between the meaning...

...since noise words aren't even listed in the index for the physical document collection. **Frequently** used terms are listed in the index. Furthermore, noise words are completely eliminated from the...

...be seen on the screen after completion of the search. Also note that if a **frequently** used term is a word, it will not be eliminated if it is part of...of importance, as shown in the column 188. At the bottom of the list are **frequently** used terms, which, as discussed above, are not part of the search. Note that the **frequently** used term "A" in column 182

shows "--" in the columns 184,186,188.
Referring to...count values. Also at the step 263, the SR(s) return an indicator for any **frequently** used terms (described above) that will be eliminated from further consideration.
Following the step 264...

...in the formula, thus weighting the ranking in favor of documents having the greatest term **frequency** .
At the step 268, each of the SR's returns to the SA the ranking...

...CLAIMS the scores are calculated using a formula that varies according to the square of the **frequency** in each of the text documents of each of the search terms, where the document **frequency** is defined as the number of documents within a searched collection which contain the search...

...system, according to claim 1, wherein the formula also varies according to an inverse document **frequency** of each of the search terms.
3. The associative text search and retrieval system, according...for the search by not including noise terms in the index; and
means for excluding **frequently** used terms from being considered for the search, the **frequently** used terms being contained in the index and maintained in a list of **frequently** used terms, the **frequently** used terms being excluded from the search by not using terms in the list for...

...each of the search terms wherein the term importance varies according to the inverse document **frequency** of the search term.
17. The associative text search and retrieval system, according to claim ...
...of the search terms using a formula that varies according to the square of the **frequency** in each of the text documents of each of the search terms (268); where the document **frequency** is defined as the number of documents within a searched collection which contain the search...

...system, according to claim 22, wherein the formula also varies according to an inverse document **frequency** of each of the search terms.
24. The method of operating an associative text search...by not including noise terms in the index;
excluding from being considered for the search **frequently** used terms contained in the index and in the document collection in a list of **frequently** used terms, the list of **frequently** used terms being dynamic, based upon a variety of functional factors including the **frequency** of occurrence of a term in the document collection and the nature of the document collection, the **frequently** used terms being excluded from the search by not using terms in the list for...

...text documents containing at least one of the search terms except for noise terms and **frequently** used terms excluded in said excluding step.
33. The method of operating an associative text...
...each of the search terms wherein the term importance varies according to the inverse document **frequency** of the search term.
36. The method of operating an associative text search and retrieval...
...CLAIMS de points sont calcules en utilisant une formule qui varie suivant le carre de la **frequence** , dans chacun des documents de textes, de chacun des termes de recherche, la **frequence** de documents etant definie comme etant le nombre de documents, a l'interieur d'une...

...selon la revendication 1, dans lequel la formule varie egalement suivant l'inverse de la **frequence** , dans le document, de chacun des termes de recherche.

3. Systeme associatif de recherche et...chacun des termes de recherche, l'importance du terme variant suivant l'inverse de la **frequence** d'apparition, dans le document, du terme de recherche

17. Systeme associatif de recherche et...des termes de recherche en utilisant une formule qui varie suivant le carre de la **frequence** d'apparition, dans chacun des documents de textes de chacun des termes de recherche (268), la **frequence** de documents etant definie comme le nombre de documents, a l'interieur d'une collection...

...selon la revendication 22, dans lequel la formule varie egalement suivant l'inverse de la **frequence** d'apparition, dans le document, de chacun des termes de recherche.

24. Procede d'exploitation...de termes frequemment utilises etant dynamique, fondee sur une variete de facteurs fonctionnels incluant la **frequence** d'apparition d'un terme dans la collection de documents et la nature de la...

...chacun des termes de recherche, l'importance du terme variant suivant l'inverse de la **frequence** dans le document du terme de recherche.

36. Procede d'exploitation d'un systeme associatif...

17/5,K/4 (Item 4 from file: 348)

DIALOG(R)File 348:EUROPEAN PATENTS

00602441

Method for resolution of natural-language queries against full-text databases

Verfahren, um natursprachliche Abfragen von Textdatenbanken zu losen

Procede pour resoudre des demandes en langage naturel dans des bases de donnees de textes

PATENT ASSIGNEE:
CONQUEST SOFTWARE INC., (1713100), 9700 Patuxent Woods Drive, Suite 140, Columbia, Maryland MD-21046, (US), (Proprietor designated states: all)
INVENTOR:
Addison,Edwin R. Conquest Software Inc., 9700 Patuxent Woods Drive,Suite 140,, Columbia,Maryland MD-21046, (US)
Blair,Arden S. Conquest Software Inc., 9700 Patuxent Woods Drive,Suite 140,, Columbia,Maryland MD-21046, (US)
Nelson,Paul E. Conquest Software Inc., 9700 Patuxent Woods Drive,Suite 140,, Columbia,Maryland MD-21046, (US)
Schwartz,Thomas Conquest Software Inc., 9700 Patuxent Woods Drive,Suite 140, Columbia,Maryland MD-21046, (US)
LEGAL REPRESENTATIVE:
Goodman, Christopher (31122), Eric Potter Clarkson, Park View House, 58 The Ropewalk, Nottingham NG1 5DD, (GB)

PATENT (CC, No, Kind, Date): EP 597630 A1 940518 (Basic)
EP 597630 B1 020731
APPLICATION (CC, No, Date): EP 93308829 931104;
PRIORITY (CC, No, Date): US 970718 921104
DESIGNATED STATES: AT; BE; CH; DE; DK; ES; FR; GB; GR; IE; IT; LI; LU; MC; NL; PT; SE
INTERNATIONAL PATENT CLASS: G06F-017/27; G06F-017/30
CITED PATENTS (EP B): EP 494573 A; US 4849898 A; US 5056021 A

ABSTRACT EP 597630 A1

The method of the present invention combines concept searching, document ranking, high speed and efficiency, browsing capabilities, "intelligent" hypertext, document routing, and summarization (machine abstracting) in an easy-to-use implementation. The method of the present invention also offers Boolean and statistical query options. The method of the present invention is based upon "concept indexing" (an index of "word senses" rather than just words.) It builds its concept index from a "semantic network" of word relationships with word definitions drawn from one or more standard human-language dictionaries. During query, users may select the meaning of a word from the dictionary during query construction, or may allow the method to disambiguate words based on semantic and statistical evidence of meaning. This results in a measurable improvement in precision and recall. Results of searching are retrieved and displayed in ranked order. The ranking process is more sophisticated than prior art systems providing ranking because it takes linguistics and concepts, as well as statistics into account. (see image in original document)

ABSTRACT WORD COUNT: 168

NOTE:

Figure number on first page: 1

LEGAL STATUS (Type, Pub Date, Kind, Text):

Change:	001129	A1	International Patent **Classification** changed: 20001013
Application:	940518	A1	Published application (A1with Search Report ;A2without Search Report)
Lapse:	040922	B1	Date of lapse of European Patent in a contracting state (Country, date): AT 20020731, BE 20020731, CH 20020731, LI 20020731, DE 20021101, DK 20021031, ES 20030130, FR 20030221, GB 20021104, GR 20020731, IE 20021104, LU 20021104, NL 20020731, PT 20021112, SE 20021031,
Lapse:	040121	B1	Date of lapse of European Patent in a contracting state (Country, date): AT 20020731, BE 20020731, CH 20020731, LI 20020731, DE 20021101, DK 20021031, ES 20030130, FR 20030221, GR 20020731, IE 20021104, NL 20020731, PT 20021112, SE 20021031,
Lapse:	031126	B1	Date of lapse of European Patent in a contracting state (Country, date): AT 20020731, BE 20020731, CH 20020731, LI 20020731, DE 20021101, DK 20021031, FR 20030221, GR 20020731, NL 20020731, PT 20021112, SE 20021031,
Lapse:	031105	B1	Date of lapse of European Patent in a contracting state (Country, date): AT 20020731, CH 20020731, LI 20020731, DE 20021101, DK 20021031, GR 20020731, NL 20020731, PT 20021112, SE 20021031,
Oppn None:	030723	B1	No opposition filed: 20030506
Lapse:	030514	B1	Date of lapse of European Patent in a contracting state (Country, date): GR 20020731, NL 20020731, SE 20021031,
Lapse:	030115	B1	Date of lapse of European Patent in a contracting state (Country, date): SE 20021031,
Change:	001213	A1	International Patent **Classification** changed: 20001026

Grant: 020731 B1 Granted patent
Lapse: 030219 B1 Date of lapse of European Patent in a contracting state (Country, date): NL 20020731, SE 20021031,
Lapse: 030528 B1 Date of lapse of European Patent in a contracting state (Country, date): GR 20020731, NL 20020731, PT 20021112, SE 20021031,
Lapse: 030730 B1 Date of lapse of European Patent in a contracting state (Country, date): AT 20020731, CH 20020731, LI 20020731, GR 20020731, NL 20020731, PT 20021112, SE 20021031,
Lapse: 031119 B1 Date of lapse of European Patent in a contracting state (Country, date): AT 20020731, CH 20020731, LI 20020731, DE 20021101, DK 20021031, FR 20030221, GR 20020731, NL 20020731, PT 20021112, SE 20021031,
Lapse: 040102 B1 Date of lapse of European Patent in a contracting state (Country, date): AT 20020731, BE 20020731, CH 20020731, LI 20020731, DE 20021101, DK 20021031, FR 20030221, GR 20020731, IE 20021104, NL 20020731, PT 20021112, SE 20021031,
Lapse: 040526 B1 Date of lapse of European Patent in a contracting state (Country, date): AT 20020731, BE 20020731, CH 20020731, LI 20020731, DE 20021101, DK 20021031, ES 20030130, FR 20030221, GB 20021104, GR 20020731, IE 20021104, NL 20020731, PT 20021112, SE 20021031,
Examination: 981111 A1 Date of despatch of first examination report: 980928
LANGUAGE (Publication,Procedural,Application): English; English; English
FULLTEXT AVAILABILITY:

Available Text	Language	Update	Word Count
CLAIMS B	(English)	200231	1139
CLAIMS B	(German)	200231	1201
CLAIMS B	(French)	200231	1291
SPEC B	(English)	200231	11289
Total word count - document A			0
Total word count - document B			14920
Total word count - documents A + B			14920

LEGAL STATUS (Type, Pub Date, Kind, Text):
...International Patent **Classification** changed...

...International Patent **Classification** changed...

...SPECIFICATION This pattern recognition approach based upon a cluster technique discussed in Duda and Hart, Pattern **Classification** and Scene Analysis, John Wiley & Sons, New York 1973 has the obvious drawback that it...

...art in text retrieval that uses natural language input is the statistical techniques developed by **Gerard Salton** of Cornell University. His research system called SMART is now used in commercial applications, for...

...systems which attempt to extract contextual understanding from natural

language statements is primarily that of **Gerard Salton** (described in Automatic Text Processing, Addison-Wesley Publishing Company, 1989.) As described therein, such systems...

...and by Donna Harmon, in a recent ASIS Journal article, (ranking on a combination of **frequency** related methods). Several commercial systems use ranking but their proprietors have never disclosed the algorithms used. Fulcrum uses (among other factors) document position and **frequency** . Personal Library Software uses inverse document **frequency** , term **frequency** and collocation statistics. Verity uses "accrued evidence based on the presence of terms defined in...abstracting meanings from natural language words using a thesaurus to determine levels of abstraction and **category** of meanings for words.

Each word is analysed for its semantic content by mapping into its **category** of meanings within each of four levels of abstraction. The preferred embodiment uses Roget's Thesaurus and Index of **Classification** to determine the levels of abstraction and **category** of meanings for words. Each of a sequence of words is mapped into the various levels of abstraction, forming a file of **category** of meanings for each of the words. The common **categories** between words are determined, and these common elements are output as data indicative of the most likely **categories** of meaning in each of the levels of abstraction to indicate the proper meaning intended...if there is still no word candidate, proper noun identification, acronym testing or a word **category** inference algorithm is activated using special rules to identify unknown words. For example, it may...

...idiom" in a rather loose sense, meaning any string of more than one word that **frequently** occurs together and implies a single meaning or concept.

6) Fuzzy spell corrector: When all...sets described above to find document using a query from the user. First, the least **frequent** word sense in the query (as determined by the inverted index) is checked. However, an...

...be idiomatic phrases, proper names that are more than one word long, or special token **categories** such as date patterns. The novel feature is the ability to quickly and efficiently rebundle...that will be varied in tests.

Prior art used this mechanism, also called "Inverse Document **Frequency** ", over words, not word meanings.

Query Augmentation to Improve Recall (horizontal bar) Using Syntactic and ...to a particular document, sentence, and phrase.

Step 6: Weighting by Quantity (AKA Inverse Document **Frequency**)

In information theory, the concepts which occur most often are the least useful (contain the...

...occurrences (a large number of occurrences will cause the weight to be reduced because a **frequently** occurring concept carries less information since it will be in many documents).

Step 7: Ranking...

...items will be sent.

The query by example feature of the present invention may be **classified** as a "context vector" approach. Context vector means a collection of terms used together in...there are systems that rank documents. Most of these systems rank the documents on the **frequency** of occurrence of the terms in the query. Some systems also take into account

the inverse document **frequency** of terms. Yet other systems rank on position giving higher weight to terms that appear...

...5 everywhere else).
Other factors in Pi)j)) include importance factors (word specificity, inverse document **frequency** and syntactic position) and closeness of match factors (semantic distance, syntactic order). An "untested" one... This is done by using spreading activation with the semantic network.
3) Determine the most **frequent** concepts in the document, using histograms or some other technique. This includes the concepts in...

...4) Construct the abstract by excerpting sentences from the original document. Sentences containing the most **frequent** concepts (or are closely related to the most **frequent** concepts) are used first. The abstract is simply a collection of these excerpted sentences.
There...

...CLAIMS c) comprises a ranking according to at least one of the following criteria: inverse document **frequency** ; syntactic position; part of speech; application of a predetermined concept tree; part of speech; predetermined...
...CLAIMS lequel l'etape (c) comprend un classement selon au moins l'un des criteres suivants : **frequence** en document inverse ; position syntaxique ; partie de langage ; application d'une arborescence conceptuelle ; partie de...

17/5,K/5 (Item 1 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT

00566612 **Image available**
METHOD AND SYSTEM FOR SUMMARIZING TOPICS OF DOCUMENTS BROWSED BY A USER
PROCEDE ET SYSTEME DE RECAPITULATION DE THEMES DE DOCUMENTS EXPLORES PAR UN UTILISATEUR
Patent Applicant/Assignee:
INTERNATIONAL BUSINESS MACHINES CORPORATION,
Inventor(s):
COHEN Andrew L,
MAGLIO Paul P,
BARRETT Robert C,
SHELDON Mark A,
Patent and Priority Information (Country, Number, Date):
Patent: WO 200029985 A1 20000525 (WO 0029985)
Application: WO 99US26992 19991115 (PCT/WO US9926992)
Priority Application: US 98191587 19981113
Designated States:
(Protection type is "patent" unless otherwise stated - for applications prior to 2004)
AL AM AT AU AZ BA BB BG BR BY CA CH CN CU CZ DE DK EE ES FI GB GD GE GH GM HR HU ID IL IN IS JP KE KG KP KR KZ LC LK LR LS LT LU LV MD MG MK MN MW MX NO NZ PL PT RO RU SD SE SG SI SK SL TJ TM TR TT UA UG UZ VN YU ZW GH GM KE LS MW SD SL SZ TZ UG ZW AM AZ BY KG KZ MD RU TJ TM AT BE CH CY DE DK ES FI FR GB GR IE IT LU MC NL PT SE BF BJ CF CG CI CM GA GN GW ML MR NE SN TD TG
Main International Patent Class: G06F-017/30
Publication Language: English
Fulltext Availability:
Detailed Description
Claims

Fulltext Word Count: 9395

English Abstract
The invention disclosed herein relates to cooperative computing environments (10) and information retrieval and management methods and systems. More particularly, the present invention relates to methods and systems for capturing and generating useful information about a user's access and use of data on a computer system (12), such as in the form of documents stored on remote servers, and making such useful information available to others. Documents on the computer system (10) are accessible through a plurality of different methods, such as by specifying an identifier or locator for the document, activating a hyperlink (14) in another document which points to the document, or navigating to the document through navigational commands in an application program (26) such as a browser (18). The method involves capturing information regarding each of the accessed documents in the set, the information including the method used to access the document, dividing the set of documents, labeling (30) each subset of documents with a topic (32), and making the labels (34) and documents accessed available to other users who wish to browse the same documents.

French Abstract
La presente invention concerne des environnements informatiques cooperatifs (10) ainsi que procedes et des systemes de gestion et d'extraction d'informations. Plus particulierement, cette invention concerne des procedes et des systemes de saisie et de production de donnees utiles relatives a l'acces d'un utilisateur et a l'utilisation des donnees dans un systeme d'ordinateur (12), par exemple sous la forme de documents memorises sur des serveurs a distance, permettant que ces informations utiles soient disponibles pour d'autres utilisateurs. Les documents contenus dans le systeme d'ordinateur (10) sont accessibles par une pluralite de differents procedes, par exemple en specifiant un identificateur ou un localisateur pour le document, en activant un lien hypertexte (14) dans un autre document citant ce document, ou encore en navigant a travers le document au moyen des commandes de navigation d'un programme d'application (26), tel qu'un explorateur (18). En outre, ce procede consiste d'abord a saisir les informations concernant chacun des documents explores parmi l'ensemble de documents, ainsi que les informations contenant le procede utilise pour acceder au document, a diviser ensuite l'ensemble de documents et a etiqueter (30) chaque sous-ensemble de documents selon un theme (32) pour enfin rendre les etiquettes (34) et les documents accessibles a d'autres utilisateurs desirant explorer les memes documents.

Fulltext Availability:
Detailed Description

Detailed Description
... of useful information is compiled from the various documents. During this process, users also make **frequent** use of navigational commands offered by the user's web browser program. such as the...

...particular page found.

If the desired information is not found after a while, the user **frequently** restarts the search process by jumping to a new, unrelated resource such as the original...with topics, step 58. In a simple embodiment, labeling is performed by selecting the most **frequent** word or phrase to appear in the portion of the usage trail.

1 5 Other...present invention relies on explicit representations of

expertise in the form of updated profiles and **taxonomies** of people and their respective expertise.

As explained above. the parsing of the usage trail...International ACM SIGIR Conference. Association for Computing Machinery. New York. June, 1992. Pages 318-')29. **Gerard Salton** . Introduction to Modern Information Retrieval, (McGraw-Hill, New York 1983).

After clustering is completed, the...

17/5,K/6 (Item 2 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT

00449236 **Image available**
SCOPE TESTING OF DOCUMENTS IN A SEARCH ENGINE
CONTROLE D'UNE CIBLE DE DOCUMENTS PAR UN MOTEUR D'INTERROGATION
Patent Applicant/Assignee:
MICROSOFT CORPORATION,
Inventor(s):
PELTONEN Kyle G,
RAJU Sitaram C V,
MILEWSKI Bartosz B,
Patent and Priority Information (Country, Number, Date):
Patent: WO 9839700 A2 **19980911**
Application: WO 98US4568 19980306 (PCT/WO US9804568)
Priority Application: US 97813618 19970307
Designated States:
(Protection type is "patent" unless otherwise stated - for applications prior to 2004)
DE GB JP AT BE CH DE DK ES FI FR GB GR IE IT LU MC NL PT SE
Main International Patent Class: G06F-017/30
International Patent Class: G06F-15:40
Publication Language: English
Fulltext Availability:
Detailed Description
Claims
Fulltext Word Count: 10421

English Abstract
A method and mechanism for responding to a query in a hierarchically organized system of documents and folders. In response to the query, a set of documents is retrieved based on specified criteria. Only documents in that set which match a specified scope are returned in a result set. Scope testing is performed on each of the documents in the set by obtaining a document identifier of each document, and then using that document identifier to obtain a document identifier of the parent folder thereof. The document identifier (80) of the parent folder is used as a key to a data structure, which stores flags indicative of whether parent folders are in the specified scope. If the flag for a given parent folder indicates that the parent folder is in scope, the document having that parent is returned in the result set. If the flag indicates that the current document is not in scope, that document is not returned. If there was not an entry in the data structure for that key, prefix matching is performed on the parent folder to determine whether it is in scope. The parent folder scope information is then added to the data structure as a flag indexed by the document identifier of the parent folder.

French Abstract
Procede et mecanisme permettant de repondre a une interrogation dans un

systeme de documents et de dossiers organise hierarchiquement. En reponse a l'interrogation, un ensemble de documents est extrait sur la base de criteres specifies. Seuls les documents se trouvant dans cet ensemble et correspondant a une cible specifiee sont retournes dans un ensemble resultat. Le controle de la cible est effectue sur chaque document de l'ensemble grace a l'obtention d'un identificateur de documents de chaque document et ensuite, par utilisation de cet identificateur de documents pour obtenir un identificateur de documents du dossier parent de ce document. L'identificateur de documents du dossier parent est utilise comme une cle a une structure de donnees qui enregistre des drapeaux qui indiquent si les dossiers parents se trouvent dans la cible specifiee. Si le drapeau, pour un dossier parent donne, indique que le dossier donne se trouve dans la cible, le document ayant ce parent est renvoye dans le jeu resultat. Si le drapeau indique que le document actuel ne se situe pas dans la cible, ce document n'est pas renvoye. S'il n'existait pas d'entree dans la structure de donnees pour cette cle, la correspondance prefixee est effectuee sur le dossier parent en vue de determiner si il se trouve dans la cible. Les informations relatives a la presence ou a l'absence du dossier parent dans la cible sont alors ajoutees a la structure de donnees sous forme d'un drapeau indexe par l'identificateur de documents du dossier parent.

Patent and Priority Information (Country, Number, Date):
  Patent: ... **19980911**
Fulltext Availability:
  Detailed Description
Publication Year: **1998**

Detailed Description
... files,
  objects or the like) are capable of being organized under folders (i.e., directories, **catalogs** or the like), wherein each document may be under a hierarchical arrangement of folders and...in "Automatic Text Processing - The Transformation Analysis and Retrieval of Information By Computer," authored by **Gerard Salton** , Addison-Wesley (1989).

  In any event, the operation of the criteria matching component 68 is...per manently stored for each available scope. Likewise, if one or more particular scopes are **frequently** selected by users, one or more caches may be stored for each of these most- **frequently** selected scopes. Thus, in the case wherein a result set or a cache is persistently...

17/5,K/7 (Item 3 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT

00449233 **Image available**
SYSTEM AND METHOD FOR ACCESSING HETEROGENEOUS DATABASES
SYSTEME ET PROCEDE D'ACCES A DES BASES DE DONNEES HETEROGENES
Patent Applicant/Assignee:
  AT & T CORP,
Inventor(s):
  COHEN William W,
Patent and Priority Information (Country, Number, Date):
  Patent: WO 9839697 A2 **19980911**
  Application: WO 98US3627 19980225 (PCT/WO US9803627)

Priority Application: US 9739576 19970225; US 9828471 19980224
Designated States:
(Protection type is "patent" unless otherwise stated - for applications prior to 2004)
CA AT BE CH DE DK ES FI FR GB GR IE IT LU MC NL PT SE
Main International Patent Class: G06F-017/00
International Patent Class: G06F-17:30
Publication Language: English
Fulltext Availability:
Detailed Description
Claims
Fulltext Word Count: 12510

English Abstract
Answers are sought for queries concerning information stored in a set of collections. The queries are done with a search server (201) which uses a network (206) to communicate with database servers (203, 204, 205) having potentially heterogeneous databases (207, 208, 209, 210, 211, 212). Each collection includes a structured entity which includes a field. A query is received that specifies a subset of the set of collections and a logical constraint between fields that includes a requirement that a first field matches a second field. The probability that the first field matches the second field is determined automatically based upon the contents of the fields. A collection of lists is generated in response to the query. Each list includes members of the subset of collections specified in the query. Each list has an estimate of the probability that the members of the list satisfies the logical constraint specified in the query.

French Abstract
L'invention concerne un systeme et un procede permettant de repondre a des requetes concernant des informations mises en memoire dans un ensemble de collections. Chaque collection comprend une entite structuree, et chaque entite structuree comprend un champ. Une requete recue specifie un sous-ensemble de l'ensemble de collections et une contrainte logique entre champs comportant une demande de correspondance entre un premier champ et un deuxieme champ. La probabilite de correspondance entre le premier champ et le deuxieme champ est determinee automatiquement sur la base du contenu des champs. Une collection de listes est generee en reponse a la requete, chaque liste comprenant des elements du sous-ensemble de collections specifie dans la requete, et chaque liste presentant une estimation de la probabilite que les elements de la liste satisfassent la contrainte logique specifiee dans la requete.

Patent and Priority Information (Country, Number, Date):
Patent: ... **19980911**
Fulltext Availability:
Detailed Description
Publication Year: **1998**

Detailed Description
... Such techniques are described in Chapters 8 and 9 of Automatic Text Processing, edited by **Gerard Salton** , Addison Wesley, Reading, Massachusetts, 1989, and ...name constants are co-referent is far from trivial in many real-world data sources. **Frequently** it requires detailed knowledge of the world, the purpose of the user's query, or... represented by v, and otherwise the value @'= (log(TFv) + 1) - log(IDFt), where the "term **frequency** " is the number of times that term t occurs in the document represented by v, and the inverse document **frequency** IDF, is IC 1 . where C, is the subset of documents in C that

C...

...known. The vector space representation for documents is described in Automatic Text Processing, edited by **Gerard Salton** , Addison Welsley, Reading, Massachusetts, 1989.

5 The general idea behind this scheme is that the...

...in the collection C. The weighting scheme also gives higher weights to terms that occur **frequently** in a document. However, in this context, this heuristic is probably not that important, since...Uniquely appearing terms like "Lucent" and "Microsoft" would have high weights. And terms of intermediate **frequency** like Acme and American would have intermediate weights.

The present invention operates on data is...the number of non-zero similarities can be greatly reduced by discarding a few very **frequent** terms like "Inc." However, even after this preprocessing, there are more than 19,000 non...

...times the number of correct pairings. This is due to a large number of moderately **frequently** terms (like "American" and "Airlines") that cannot be safely discarded. Thus, it is in general...that insiderTip contains the vector xi, corresponding to the document "Armadillos, Inc." Due to the **frequent** occurrence of the term "Inc.", there will be many documents Y that have non-zero...

...somewhat similar to xi. Recall that the weight of a term depends inversely on its **frequency** , so rare terms have high weight, and hence these Y's will share at least...

...most similar to the document "The American Software Company" (in which every term is somewhat **frequent**), a very different type of subplan might be required. The observations suggest that query processing...match at most only four terms: "services" "and", "or", "equipment", all of which are relatively **frequent** , and hence have low weight.
Next, a state will again be removed from the OPEN...used to bind Company2 and Website. Note that bindings are generated using high-weight, low-**frequency** terms first, and low-weight, high- **frequency** terms only when necessary.

Embodiments of the invention have been evaluated on data collected from ...

17/5,K/8 (Item 4 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT

00437033 **Image available**
BROWSER FOR USE IN NAVIGATING A BODY OF INFORMATION, WITH PARTICULAR APPLICATION TO BROWSING INFORMATION REPRESENTED BY AUDIOVISUAL DATA
LOGICIEL D'EXPLORATION AUX FINS DE LA NAVIGATION DANS UN BLOC D'INFORMATIONS, NOTAMMENT AUX FINS DE L'EXPLORATION D'INFORMATION SOUS FORME DE DONNEES AUDIOVISUELLES
Patent Applicant/Assignee:
INTERVAL RESEARCH CORPORATION,
Inventor(s):
AHMAD Subutai,

BHADKAMKAR Neal A,
COUSINS Steve B,
FARBER Emanuel E,
FREIBERGER Paul A,
HORNER Christopher D,
PIERNOT Philippe P,
ULLMER Brygg A,
Patent and Priority Information (Country, Number, Date):
Patent: WO 9827497 A1 **19980625**
Application: WO 97US22145 19971203 (PCT/WO US9722145)
Priority Application: US 96761030 19961205
Designated States:
(Protection type is "patent" unless otherwise stated - for applications prior to 2004)
AL AM AT AU AZ BA BB BG BR BY CA CH CN CU CZ DE DK EE ES FI GB GE GH HU ID IL IS JP KE KG KP KR KZ LC LK LR LS LT LU LV MD MG MK MN MW MX NO NZ PL PT RO RU SD SE SG SI SK SL TJ TM TR TT UA UG UZ VN YU ZW GH KE LS MW SD SZ UG ZW AM AZ BY KG KZ MD RU TJ TM AT BE CH DE DK ES FI FR GB GR IE IT LU MC NL PT SE BF BJ CF CG CI CM GA GN ML MR NE SN TD TG
Main International Patent Class: G06F-017/30
Publication Language: English
Fulltext Availability:
Detailed Description
Claims
Fulltext Word Count: 27355

English Abstract
The invention facilitates and enhances review of a body of information (that can be represented by a set of audio data, video data, text data or some combination of the three), enabling the body of information to be quickly reviewed to obtain an overview of the content of the body of information and allowing flexibility in the manner in which the body of information is reviewed. In a particular application of the invention, the content of audiovisual news programs is acquired from a first set of one or more information sources (e.g., television news programs) and text news stories are acquired from a second set of one or more information sources (e.g., on-line news services or news wire services). In such a particular application, the invention can enable the user to access the news stories of audiovisual news programs in a random manner so that the user can move quickly among news stories or news programs. The invention can also enable the user to quickly locate news stories pertaining to a particular subject. Additionally, when the user is observing a particular news story in a news program, the invention can identify and display related news stories. The invention can also enable the user to control the display of the news programs by, for example, speeding up the display, causing a summary of one or more news stories to be displayed, or pausing the display of the news stories. Additionally, the invention can indicate to the user which news story is currently being viewed, as well as which news stories have previously been viewed.

French Abstract
L'examen d'un bloc d'information effectue grace aux techniques de cette invention est facilite et ameliore, ces informations pouvant etre des donnees audio ou video, des donnees textuelles ou bien une combinaison des trois types. Cette invention, qui permet de passer rapidement en revue le bloc d'information, renforce la souplesse de la procedure d'examen. Dans un mode de realisation, le contenu d'emissions de nouvelles audiovisuelles est acquis a partir d'un premier jeu de sources d'information (des programmes televises, par exemple) et celui de magazines de nouvelles textuelles a partir d'un second jeu de sources d'information (des journaux electroniques ou des services cables, par

exemple). Dans ce mode de realisation, l'utilisateur a acces aux magazines de nouvelles textuelles des emissions de nouvelles audiovisuelles de maniere aleatoire, de sorte qu'il lui est possible de se deplacer rapidement dans ces programmes, textuels et audiovisuels. Cette invention permet egalement a l'utilisateur de localiser rapidement des magazines de nouvelles traitant d'un sujet particulier. De surcroit, lorsque cet utilisateur visionne un magazine particulier dans une emission de nouvelles, il est a meme, grace a cette invention, de recenser des magazines et de les afficher. Cette invention lui permet egalement d'agir sur l'affichage des emissions de nouvelles, par exemple, en accelerant la cadence d'affichage, en obtenant des sommaires du ou des magazines a afficher ou en faisant des pauses. Il lui est possible, en outre, de savoir quel est le magazine qu'il regarde actuellement et quels sont ceux qu'il a vu precedemment.

Patent and Priority Information (Country, Number, Date):
Patent: ... **19980625**
Fulltext Availability:
Detailed Description
Claims
Publication Year: **1998**

Detailed Description
... particular interest. In particular, there is a need for 10 systems and methods of organizing, **categorizing** and relating the various segments of a large body of information to facilitate the access...

...related segments have previously been determined or, at least, that the 20 segments have been **categorized** according to subject matter content so that whether two segments are related can readily be...these systems generally do not analyze the data to enable the data to be organized, **categorized** and related so that, for example, segments of the body of information can be related...

...quickly ascertain the subject matter content of the news stories contained therein. Additionally, a particular **category** (e.g., subject matter **category**) can be specified and news stories having content that fits within the specified subject matter **category** can be immediately identified and either displayed or identified as pertinent to the subject matter **category** and available for display. Further, a user of SUBSTITUTE SHEET (RULE 26) the news browser...

...single representative video frame displayed for each such news story.

Additionally, the invention enables automatic **categorization** of uncategorized segments of the body of 30 information based upon comparison to other segments of the body of information that have been **categorized** . In particular, the subject matter **category** of a segment of information can be determined by comparing the segment to one or more previously SUBSTITUTE SHEET (RULE 26) **categorized** segments and **categorizing** the segment in accordance with the subject matter **categorization** of one or more previously **categorized** segments that are determined to be

The content of the segment could be determined, for example, using a **categorization** method as described in more detail 5 below. The segment to be **categorized** could either be compared to previously **categorized** segments that can be displayed by the system of the invention, or to a library...for example, by the data storage
device 104) . Data can be acquired at any desired **frequency** and 25 the scheduled acquisition times specified in any desired manner (e. g. , hourly , daily...Thus, it is
necessary or desirable to "structure" the data (i.e. , to organize and **categorize** the data, and relate particular data to other data) in useful ways. Below are described...correlation of primary 15 information segments with secondary information segments can also be used to **categorize** the primary information segments according to subject matter, thus enabling the user to sort or...

...cause display of segments of the primary information that pertain to a particular subject matter **category** (see the 20 discussion of the topic buttons 215 ...text transcripts of 5 bodies of information represented as a set of audiovisual information also **frequently** include markers that identify breaks between segments of the information. For example, closed caption text...and video data is obtained. However, synchronization
between the text data and the audio data **frequently** does not already exist, and, if it does not, obtaining such 30 synchronization can be...

...observed that the video data of a news story from an 10 audiovisual news program **frequently** begins about 5 to 10 seconds before the closed caption text data of the news...the determination of relatedness between segments of information represented by audiovisual data (such as is **frequently** the case for the primary information that can be displayed by the invention) and segments...

...display of the related secondary information region 204 to be generated. It can also enable **categorization** of uncategorized segments, as described further below.

FIG. 4 is a flow chart of a...in more detail in, for example, the textbook entitled
Introduction to Modern Information Retrieval, by **Gerard Salton** , McGraw-Hill, New York, 1983, the pertinent disclosure of which is incorporated by reference herein...that
reveal the situation described above before discarding any of the secondary information segments.

3. **Categorizing**
An important aspect of the invention is the capability to 15 **categorize** uncategorized segments of information based ill-,on the **categorization** of previously **categorized** segner t s f information. In particular, if the segments of the secondary information have been' **categorized** according to subject matter, then the degree of similarity between the subject matter 20 content...

...secondary information (e.g. news stories from text news

In step 5 1 t e uncategorized segment is **categorized**
20 based upon the subject matter **categories** associated with the relevant previously **categorized** segments. One or more subject matter **categories** can be associated with the uncategorized segment. Generally, the subject matter **category** or **categories** can be selected from the subject matter **categories** associated
25 with the relevant previously **categorized** segments using any desired method. For example, the subject matter **category** or **categories** of the most similar previously **categorized** segment could be selected as the subject matter **category** or **categories** of the uncategorized segment. Or, the most **frequently**
30 occurring subject matter **category** or **categories** associated with a predefined number of the most similar previously **categorized** segments (or previously **categorized** segments having greater than a threshold degree of similarity) could be selected as the
SUBSTITUTE SHEET (RULE 26)
subject matter **category** of the uncategorized segment. In the latter case, it may be particularly desirable, as described above, to determine the similarity between the relevant previously **categorized** segments, so that only one of a set of
5 previously **categorized** segments that are substantially identical to each other influences the **categorization** of the uncategorized segment.
C. Information Presentation
Above, the acquisition of information and the structuring...text data is obtained, the text data can be "pre-processed" using known methods to **classify** the words in the text data according to
5 their characteristics, e.g., part of...

Claim
... region includes an interface
that enables selection of one of a plurality of subject matter **categories** ; and
the means for controlling further comprises:
means for identifying the subject matter **category** of a segment; and
means for controlling the system to display each of the segments that correspond to a selected subject matter **category** .

23 A system as in Claim 21, wherein:
the playback control region includes an
that...a first segment to which the second segment is related.
3 6 .A method for **categorizing** according to subject matter
15 an uncategorized segment of a body of information that includes...

...of information, one or more
segments of the body of information having previously Le-en **categorized** by identifying each of the one or more segment's
20 with one or more subject matter **categories** , the method comprising the steps of:
determining the degree of similarity betvATeen
subject matter content of the uncategorized secrment and
the subject matter content of each of the previously **categorized** segments;
identifying one or more of the previously **categorized** segments as relevant to the uncategorized segment based upon the determined degrees of similarity of subject matter content between the uncategorized segment and the previously **categorized** segments; and

uncategorized segment and the previously **categorized** segments; and
instructions for selecting one or more subject matter **categories** with which to identify the uncategorized segment based upon the subject matter **categories** t--, identify the relevant previously **categorized** segments.

61 A computer readable medium encoded with one or more computer programs for enabling...

17/5,K/9 (Item 5 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT

00406217 **Image available**
FINDING AN E-MAIL MESSAGE TO WHICH ANOTHER E-MAIL MESSAGE IS A RESPONSE
RECHERCHE D'UN MESSAGE ELECTRONIQUE AUQUEL REPOND UN AUTRE MESSAGE ELECTRONIQUE
Patent Applicant/Assignee:
AT & T CORP,
Inventor(s):
KNOWLES Kimberly A,
LEWIS David Dolan,
Patent and Priority Information (Country, Number, Date):
Patent: WO 9746962 A1 **19971211**
Application: WO 97US9161 19970530 (PCT/WO US9709161)
Priority Application: US 9619264 19960607; US 97866196 19970530
Designated States:
(Protection type is "patent" unless otherwise stated - for applications prior to 2004)
CA JP MX AT BE CH DE DK ES FI FR GB GR IE IT LU MC NL PT SE
Main International Patent Class: G06F-017/60
International Patent Class: G06F-17:30; G06F-17:20; H04L-12:58
Publication Language: English
Fulltext Availability:
Detailed Description
Claims
Fulltext Word Count: 27085

English Abstract
Current tools for processing e-mail and other messages do not adequately recognize and manipulate threads, i.e., conversations among two or more people carried out by exchange of messages. The present invention utilizes the textual context and characteristics of messages in order to provide a more reliable and effective way to construct message threads. In accordance with the present invention, statistical information retrieval techniques are used in conjuction with textual material obtained by "filtering" of messages to achieve a significant level of accuracy at identifying when one message is a reply to another.

French Abstract
Les outils actuels de traitement de messages electroniques et autres messages ne reconnaissent pas et ne manipulent pas de facon adequate les enchainements, c'est-a-dire les conversations entre deux ou plusieurs personnes sous la forme d'echanges de messages. La presente invention utilise le contexte et les caracteristiques textuels des messages afin de reconstituer des enchainements de messages de facon plus fiable et efficace. Selon cette invention, des techniques statistiques de recherche d'informations sont utilisees conjointement avec des documents textuels

obtenus par "filtrage" des messages afin de garantir un niveau de precision important lors de l'identification, lorsqu'un message repond a un autre message.

Patent and Priority Information (Country, Number, Date):
Patent: ... **19971211**
Fulltext Availability:
Detailed Description
Publication Year: **1997**

Detailed Description
... their
content is unlikely to be 100% accurate. The effectiveness 20 of content-based text **categorization** systems varies considerably among **categories** , and accuracies over 95% are rarely reported. This means that threads having as few as...on distinguishing what relationship a particular cue phrase is indicating can be applied.

(7) Message **Categorization** . Certain types of messages to such as calls for papers and job ads are unlikely...

...replies to other messages and/or are unlikely to be replied to publicly. Known text **categorization** methods can detect these and provide evidence against the presence of response links.

(8) Detection...of the SMART system, a publically available text retrieval system described in.

@book(SALTON83
author " **Gerard Salton** and Michael J. McGill"
title "Introduction to Modern Information Retrieval"
publisher = "McGraw-Hill Book Company...of complexity is that (in the limitl it arch.8906:Subject@ Kolmogorov-Chaitin complexity (was: **Categorization** and Supervision) arch.8906: about the Kolmogorov-Chaitin view of complexity is that (in the...

17/5,K/10 (Item 6 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT
(c) 2005 WIPO/Univentio. All rts. reserv.

00347148 **Image available**
RETRIEVAL OF HYPERLINKED INFORMATION RESOURCES USING HEURISTICS
EXTRACTION DE RESSOURCES D'INFORMATIONS HYPERLIEES UTILISANT DES PROCEDES HEURISTIQUES
Patent Applicant/Assignee:
INTERVAL RESEARCH CORPORATION,
Inventor(s):
SHOHAM Yoav,
Patent and Priority Information (Country, Number, Date):
Patent: WO 9629661 A1 **19960926**
Application: WO 96US2572 19960226 (PCT/WO US9602572)
Priority Application: US 95681 19950320; US 95206 19950512
Designated States:
(Protection type is "patent" unless otherwise stated - for applications prior to 2004)
AL AM AT AU AZ BB BG BR BY CA CH CN CZ DE DK EE ES FI GB GE HU IS JP KE KG KP KR KZ LK LR LS LT LU LV MD MG MK MN MW MX NO NZ PL PT RO RU SD SE

SG SI SK TJ TM TR TT UA UG UZ VN KE LS MW SD SZ UG AZ BY KG KZ MD RU TJ TM AT BE CH DE DK ES FR GB GR IE IT LU MC NL PT SE BF BJ CF CG CI CM GA GN ML MR NE SN TD TG

Main International Patent Class: G06F-017/30

Publication Language: English

Fulltext Availability:
Detailed Description
Claims

Fulltext Word Count: 8261

English Abstract

A system and method for adaptively transversing a network of linked textual or multi-media information utilizes one or more heuristics to explore the network and present information to a user. An exploration or search heuristic (124) governs activity while examining and exploring the linked information resources, while a presentation heuristic controls presentation of a manageable amount of information resources to the user (126). The system and method accept relevance feedback (128) from the user which is used to refine the future search, retrieval, and presentation of information resources (130). The user may present an information query of various degrees of specificity or the system and method may search and present information resources based entirely on relevance feedback from the user.

French Abstract

L'invention concerne un systeme et un procede d'exploration transversale en mode adaptatif d'un reseau d'informations textuelles ou multimedia liees, mettant en oeuvre une ou plusieurs techniques heuristiques (124), dans le but de presenter des informations a l'utilisateur. L'exploration ou la recherche des ressources d'information liees se font selon des techniques heuristiques, et cette methode preside aussi a la presentation a l'utilisateur (126) d'une quantite gerable de ressources d'informations. Le systeme et le procede permettent a l'utilisateur de fournir en retour une information de pertinence (128), qui sert a affiner la recherche, l'extraction et la presentation subsequentes de ressources d'informations (130). L'utilisateur peut adresser une demande d'information presentant differents degres de specificite ou bien le systeme peut rechercher et presenter des ressources d'informations fondees entierement sur l'information de pertinence fournie en retour par l'utilisateur.

Patent and Priority Information (Country, Number, Date):
Patent: ... **19960926**

Fulltext Availability:
Detailed Description

Publication Year: **1996**

Detailed Description

... collection is to add structure to the collection. For example, information is often sorted and **classified** so that a large portion of the collection need not be searched. However, this type of structure often requires some familiarity with -3 the **classification** system, to avoid elimination of relevant resources by improperly limiting the search to a particular **classification** or group of **classifications** .

Another approach ...approach utilizing standard information retrieval techniques consists of systematically exploring the network and generating a

catalog , index, or map of links associated with documents containing information of interest. This index is...

...addition, standard information retrieval techniques require the user to articulate or characterize information of interest. **Frequently** , however, users may be able to easily recognize a document meeting their information need, but...resource will be to the user. Examples of such techniques are discussed in detail by **Gerard Salton** and Michael J. McGill, An Introduction to Modern Information Retrieva7, McGraw-Hill, 1983.

In one embodiment of the present invention, a metric such as the product of term **frequency** and inverse document **frequency** (TFIDF) is used. The value of an object k (such as a word) in a...tf (i) is the number of times word di appears in document T (the term **frequency**) , df (i) is the number of documents in the collection which contain di (the document **frequency**) , n is the number of documents in the collection and tfmax is the maximum term **frequency** over all words in T.

The document **frequency** component was calculated using a f ixed dictionary of approximately 27,000 words gathered from...

17/5,K/11 (Item 7 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT

00346274
METHOD AND SYSTEM FOR TWO-DIMENSIONAL VISUALIZATION OF AN INFORMATION TAXONOMY **AND OF TEXT DOCUMENTS BASED ON TOPICAL CONTENT OF THE DOCUMENTS**
PROCEDE ET SYSTEME DE VISUALISATION BIDIMENSIONNELLE D'UNE TAXONOMIE **D'INFORMATIONS ET DE DOCUMENTS DE TEXTES EN FONCTION DU SUJET DES DOCUMENTS**
Patent Applicant/Assignee:
BARTELL Brian,
CLARKE Robert,
Inventor(s):
BARTELL Brian,
CLARKE Robert,
Patent and Priority Information (Country, Number, Date):
Patent: WO 9628787 A1 **19960919**
Application: WO 96US3411 19960312 (PCT/WO US9603411)
Priority Application: US 95402839 19950313
Designated States:
(Protection type is "patent" unless otherwise stated - for applications prior to 2004)
AL AM AT AU AZ BB BG BR BY CA CH CN CZ DE DK EE ES FI GB GE HU IS JP KE KG KP KR KZ LK LR LS LT LU LV MD MG MK MN MW MX NO NZ PL PT RO RU SD SE SG SI SK TJ TM TR TT UA UG UZ VN KE LS MW SD SZ UG AM AZ BY KG KZ MD RU TJ TM AT BE CH DE DK ES FI FR GB GR IE IT LU MC NL PT SE BF BJ CF CG CI CM GA GN ML MR NE SN TD TG
Main International Patent Class: G06F-017/30

Publication Language: English
Fulltext Availability:
  Detailed Description
  Claims
Fulltext Word Count: 7827

English Abstract
A method and system for aiding users in visualising the relatedness of retrieved text documents and the topics to which they relate comprises training a **classifier** by semantically analyzing an initial group of manually- **classified** documents, positioning the classes and documents in two-dimensional space in response to semantic associations between the classes, and displaying the classes and documents. The displayed documents may be retrieved by an information storage and retrieval subsystem in any suitable manner.

French Abstract
L'invention concerne un procede et un systeme pour permettre aux utilisateurs de visualiser la relation entre les documents de textes extraits et les sujets qu'ils concernent. Ce procede et ce systeme consistent a former un **classificateur** en analysant semantiquement un groupe initial de documents classes manuellement, puis en positionnant les classes et documents dans un espace bidimensionnel en reponse aux associations semantiques entre les classes, et enfin, en affichant ces classes et ces documents. Il est possible d'extraire les documents visualises par un sous-systeme de stockage et d'extraction des informations, de toute maniere appropriee.

METHOD AND SYSTEM FOR TWO-DIMENSIONAL VISUALIZATION OF AN INFORMATION TAXONOMY **AND OF TEXT DOCUMENTS BASED ON TOPICAL CONTENT OF THE DOCUMENTS**

PROCEDE ET SYSTEME DE VISUALISATION BIDIMENSIONNELLE D'UNE TAXONOMIE **D'INFORMATIONS ET DE DOCUMENTS DE TEXTES EN FONCTION DU SUJET DES DOCUMENTS**

Patent and Priority Information (Country, Number, Date):
  Patent: ... **19960919**
Fulltext Availability:
  Detailed Description
  Claims

English Abstract
...relatedness of retrieved text documents and the topics to which they relate comprises training a **classifier** by semantically analyzing an initial group of manually- **classified** documents, positioning the classes and documents in two-dimensional space in response to semantic associations...

French Abstract
...et les sujets qu'ils concernent. Ce procede et ce systeme consistent a former un **classificateur** en analysant semantiquement un groupe initial de documents classes manuellement, puis en positionnant les classes...
Publication Year: **1996**

Detailed Description
METHOD AND SYSTEM FOR TWO-DIMENSIONAL VISUALIZATION
OF AN INFORMATION **TAXONOMY** AND OF TEXT DOCUMENTS
BASED ON TOPICAL CONTENT OF THE DOCUMENTS
BACKGROUND OF THE INVENTION...

...terms. Alternatively, a user may select a topic, and the system searches

for all documents **classified** under that topic.

Topics may be arranged in accordance with a predetermined hierarchical **classification** system. Regardless of the entry path, the system may locate
many documents, some of which...

...further refine the
search. For a comprehensive treatment of relevance ranking and relevance feedback, see **Gerard Salton** , editor, The Smart Retrieval System - Experiments in Automatic Document Processin , NJ, Prentice Hall, 1971; **Gerard Salton** , "Automatic term class construction using relevance -- a
summary of work in automatic pseudoclassification, " Information Processing & Mana-gement, 1 6:1-1 5, 1 980; **Gerard Salton** et al., Introduction to Modern Information Retrieval, McGraw-Hill, 1983.

Practitioners in the art have...The system determines the relatedness between a document and a POI in response to the **frequency** with which the keywords corresponding to the POI occur in the document.

The system thus...

...the screen and tokens representing less similar documents farther apart.

Systems are known that automatically **classify** documents in an information retrieval system under a predetermined set of classes or a predetermined hierarchical **taxonomy** to aid searching. The objective in text **classification** is to analyze an arbitrary document and determine its topical content with respect to a...

...a typical system, a computer executes an algorithm that statistically analyzes a set of manually **classified** documents, i.e., documents that have been **classified** by a human, and uses the resulting statistics to build a characterization of
"typical" documents for a class. Then, the system **classifies** each new document to be stored in the system, i.e, an arbitrary document that has not been previously **classified** , by determining the statistical similarity of the document to the prototype. Text **classification** methods include nearestneighbor **classification** and Bayesian **classification** in which the features of the Bayesian **classifier** are the occurrence of terms in the documents.

It would be desirable to simultaneously visualize...

...THE INVENTION
The present invention comprises an electronic document storage and retrieval subsystem, a document **classifier** , and a visualization subsystem.

The present invention aids users in visualizing the relatedness of retrieved...

...the topics to which they relate.

5 Documents stored in the retrieval subsystem are manually **classified** , i.e., **categorized** by human operators or editors, into a predetermined set of
classes or topics. An automatic **classifier** is constructed by calculating

frequency in the selected class (c), is initialized to zero. At step 74, the first term...

...and the method returns to step 72. Step 42 (Fig. 4) of 5 generating term **frequency** statistics is completed when all classes have been processed.

Referring again to Fig. 4, step...

...documents of each pair of classes are semantically related. Classes that use terms in similar **frequency** are more semantically related than those with dissimilar **frequencies** . Step 86 uses the term **frequency** statistics to determine the semantic association between each pair of classes. As described below, the term **frequency** statistics define class conditional probability distributions. The class conditional probability of a term, (F,,c...class conditional probabilities of the term (t).

At step 98, it is determined whether the **frequency** of the term (t) in both classes is zero. If the **frequency** is zero, the chi-squared measure is ignored, and the method proceeds to step 100...

...an embodiment of the present invention in which the classes are arranged in a flat **taxonomy** , step 1 08 comprises the steps illustrated in Fig 7. At step 1 1 0...an embodiment of the present invention in which the classes are arranged in a hierarchical **taxonomy** , step 1 08 comprises the steps illustrated in Fig. 8. Figure 8 is particularly arranged...populating the semantic space map with documents in the database comprises the step 200 of **classifying** the documents and the step 202 of positioning the documents in the semantic space map. A maximum likelihood estimation determines the most probable class for a document using the term **frequency** statistics generated at step 42. The database may include new documents 204 as well as the manually- **classified** initial documents 40.

Step 200 comprises the steps illustrated in Fig. 9. The method **classifies** a document (d) in a class (c) by estimating the semantic association between the document...is selected. Bayesian classifiers require a non-zero weight for each term. Therefore, if the **frequency** (F,,,,) of the term t is zero 1 5 in class c, it is replaced with a predetermined constant K that is small in relation to the **frequency** of the remaining terms. At step 21 2, it is determined whether Ft,c is zero. If F,,c is zero, at step 21 4, a modified **frequency** @ is set equal to the constant K, which is preferably between about 0.001 and...

...range for constant K was empirically estimated based on cross-validation experiments in which the **classifier** was trained using a portion of the documents in the database and then used to **classify** the remaining documents. This range for constant K yielded acceptable **classification** results. If Ft,c is non-zero, at step 21 6, the modified **frequency** @c,d is set equal to F,,c. At step 21 8, the natural logarithm...

...step 226, the next class c is selected,

and the method returns to step 208. **Classification** is complete when all classes c have been processed. The result of **classification** step 200 (Fig.

4) is a set of class scores Pc,d for document d.

The step 202 (Fig. 4) of positioning a **classified** document d in the semantic space map comprises the steps illustrated in Fig. 1 0...all documents, including both new
documents 204 added to the database and the initial manually- **classified** documents 40, as indicated by the dashed line in Fig. 4.

As illustrated in Figure...

Claim
...said step of generating
a semantic space map further comprises the step of generating term **frequency** statistics in response to said plurality of documents.

4 The method claimed in claim 3...

...producing a set of class scores for each said document using a 1 0 statistical **classifier** to produce a **classified** document in response to said
terms in said document; and
positioning said **classified** document in said plurality of dimensions.

7 The method claimed in claim 6, wherein:
1 5 said statistical **classifier** is trained in response to said plurality of
documents; and
said step of populating said...

...steps
of producing a set of class scores for a new document to produce a **classified** new document, and positioning said **classified** new document in
said plurality of dimensions.

8 The method claimed in claim 6, wherein...said
class and every other said class comprises the step of computing the probability and **frequency** of each said term occurring in each said class; and
said set of class scores...

...summing, over said terms in each said document, the
logarithm of the quotient of the **frequency** of said term occurring in said class divided by the sum of the sum of the **frequencies** of all terms occurring
in said class; and
if the probability of a term occurring...

...the
logarithm of the quotient of a predetermined constant divided by the sum of the **frequencies** of all terms occurring in said class.
1 1. The method claimed in claim 1...
...0
1 2. The method claimed in claim 6, wherein said step of positioning said **classified** document comprises the steps of:
selecting a predetermined number of class scores in each set to said user

query;
a **classification** subsystem for computing a semantic relatedness between each class and every other one of said...

...A machine-readable computer data storage medium having stored therein a program, comprising:
a term **frequency** statistics generator for generating term **frequency** statistics in response to a plurality of pre- **classified** documents, each having
a plurality of terms and each associated with a predetermined one of...

...classes and every other class of said plurality of classes in response to said term **frequency** statistics;
a multidimensional scaler for positioning said plurality of classes in a plurality of dimensions...

...to said semantic association between each said class and every other said class;
a statistical **classifier** for producing a set of class scores for a document in response to **frequencies** of terms in said document and for positioning said document in said semantic space map...

...probability distribution.
1 9. The data storage medium claimed in claim 16, wherein said statistical **classifier** computes said set of class scores in response to a maximum likelihood estimation.

20 The...

...wherein:
if the probability of a term occurring in a class is nonzero, said statistical **classifier** computes each class score by summing, over said terms in each said pre- **classified** document, the logarithm of the quotient of the **frequency** of said term occurring in said class divided by the sum of the sum
of the **frequencies** of all terms occurring in said class; ...and
if the probability of a term occurring in a class is zero, said statistical
classifier computes each class score by summing, over said terms in each
said class, the logarithm of the quotient of a predetermined constant divided by the sum of the **frequencies** of all terms occurring in said class.

21 The data storage medium claimed in claim...

17/5,K/12 (Item 8 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT

00296822 **Image available**
ASSOCIATIVE TEXT SEARCH AND RETRIEVAL SYSTEM
SYSTEME ASSOCIATIF DE RECHERCHE ET DE RECUPERATION DE TEXTE
Patent Applicant/Assignee:
THE MEAD CORPORATION,
Inventor(s):
HOLT John,
MILLER David James,
LU Allan X,

DALEY Ray,
DOAN Minh,
GRAHAM Richard G,
LEININGER Catherine,
McBEATH Darin W,
PEASE Thomas,
SEVER Stephen M,
WADDELL Dale,
WECKESSER Franz,
Patent and Priority Information (Country, Number, Date):
Patent: WO 9514973 A1 **19950601**
Application: WO 94US13272 19941122 (PCT/WO US9413272)
Priority Application: US 93155304 19931122
Designated States:
(Protection type is "patent" unless otherwise stated - for applications prior to 2004)
CA JP AT BE CH DE DK ES FR GB GR IE IT LU MC NL PT SE
Main International Patent Class: G06F-017/30
Publication Language: English
Fulltext Availability:
Detailed Description
Claims
Fulltext Word Count: 12236

English Abstract

An associative text search and retrieval system (30) uses one or more front end processors (56-58) to interact with a network (62) having one or more user terminals (64-66) connected thereto to allow a user to provide information to the system (30) and receive information from the system. The system (30) also includes storage (46-49) for a plurality of text documents, and at least one processor (42-44), coupled to the front end processors (56-58) and the document storage (46-49). Each of the processors (42-44) is provided access to thesaurus dictionaries (52-54). The processor(s) (32-35) search the text documents according to a search request provided by the user and provide to the front end processor (56-58) a predetermined number of retrieved documents containing at least one term of the search request. The retrieved documents have higher ranks then documents not provided to the front end processor (56-58). The ranks are calculated using a formula that varies according to the square of the **frequency** in each of the text documents of each of the search terms.

French Abstract

Systeme (30) associatif de recherche et de recuperation de texte dans lequel un ou plusieurs processeurs (56-58) frontaux servent a interagir avec un reseau (62) comprenant un ou plusieurs terminaux (64-66) d'utilisateurs connectes a ce dernier afin de permettre a un utilisateur de fournir des informations au systeme (30) et d'en recevoir de ce dernier. Ce systeme (30) comprend egalement des moyens de stockage (46-49) destines a stocker une pluralite de documents textuels, et au moins un processeur (42-44) qui est couple aux processeurs frontaux (56-58) et aux moyens de stockage (46-49) de documents. Chaque processeur (42-44) a acces a des dictionnaires thesaurus (52-54). Le ou les processeurs (32-35) recherchent les documents textuels suivant une demande de recherche formulee par l'utilisateur et fournissent au processeur frontal (56-58) un nombre predetermine de documents retrouves qui contiennent au moins un terme de la demande de recherche. Les documents retrouves ont des valeurs de classement superieures a celles des documents non fournis au processeur frontal (56-58). Les valeurs de classement sont calculees a l'aide d'une formule qui varie en fonction du carre de la **frequence** dans chacun des documents textuels de chacun des termes de recherche.

Patent and Priority Information (Country, Number, Date):
Patent: ... **19950601**
Fulltext Availability:
Detailed Description
Claims

English Abstract
...The ranks are calculated using a formula that varies according to the square of the **frequency** in each of the text documents of each of the search terms.

French Abstract
...calculees a l'aide d'une formule qui varie en fonction du carre de la **frequence** dans chacun des documents textuels de chacun des termes de recherche.
Publication Year: **1995**

Detailed Description
... phrase or multiple Associative retrieval, a technique for information retrieval developed in the 1960s by **Gerard Salton** , addresses some of the shortcomings of Boolean searching.

Automatic Text Processing, (published by Addison Wesley, New York, New York 1988, and written by **Gerard Salton**) provides a description of associative retrieval searching.

The basic formula used in associative retrieval involves...

...weighting factors are known as the terinfirequency - tf -and the inverse documentftequency - idf.

The term **frequency** is defined as the number of times the term occurs within a given document. Hence, the term **frequency** must be calculated for each document within the collection.

The inverse document **frequency** is defined as the inverse of the number of documents in the entire collection which...
...each document using a ranking formula that varies according to the square of the term **frequency** of each of the search terms in the document. The ranking formula can also vary according to the inverse document **frequency** of each search term. The formula can also use a maximum term **frequency** to estimate the size of a document and the maximum document **frequency** to estimate the number of documents in a collection of documents, thus reducing the amount...

...distinguishing between noise words, which are not provided in an index for the documents, and **frequently** used terms, which are provided in the index but which are not used in the...

...indicates the importance of each term, which varies according to the inverse of the document **frequency** of each term. The documents can be sorted according to rank or according to a...two terms 173 "FIRST" and "CASE".

The asterisk indicates that the terms which follow are " **frequently** used terms". Any terms that are deemed **frequently** used terms are eliminated from further processing for the search because their value in locating...

...search illustrated by the screen 170, the two terms 173 "FIRST" and

"CASE" were deemed **frequently** used terms and hence were not searched. The ancillary data 76, shown in connection with FIG. 3, contains a table of **frequently** used terms for each physical document collection. The determination as to which terms go into...

...known to one of ordinary skill in the art, including, but not limited to, the **frequency** of occurrence of a term in the physical collection and the relationship between the meaning...

...since noise words aren't even listed in the index for the physical document collection. **Frequently** used terms are listed in the index. Furthermore, noise WO 95/14973 PCr/US94/13272...of importance, as shown in the column 188. At the bottom of the list are **frequently** used terms, which, as discussed above, are not part of the search. Note that the **frequently** used term "A" in column 182 shows "--" in the columns 184,186,188.

Referring to...count values. Also at the step 263, the SR(s) return an indicator for any **frequently** used terms (described above) that will be eliminated from further consideration.

Following the step 264...in the formula, thus weighting the ranking in favor of documents having the greatest term **frequency** ,
At the step 268, each of the SR's returns to the SA the ranking...

Claim
... wherein said processor ranks said documents according to the square of the log of the **frequency** of occurrence of search terms contained in the documents.
Claims

2 An associative text search...

...the ranks are calculated using a formula that varies according to the square of the **frequency** in each of the text documents of each of the search terms.

3 An associative...

...system, according to claim 2, wherein the formula also varies accordina to an inverse document **frequency** of each of the search terms.

4 An associative text search and retrieval system, according...for the search by not
including noise terms in the index; and
means for excluding **frequently** used terms from being considered for the search, the **frequently** used terms being contained in the index and maintained in a list of **frequently** used terms, the **frequently** used terms being excluded from the search by not using terms in the list for
...

...for each of the search terms wherein the term importance varies according to inverse document **frequency** of the search term.

11 An associative text search and retrieval system, according to claim...
for the search by not
including noise terms in the index; and
means for excluding **frequently** used terms from being considered for the search, the **frequently** used terms being contained in the index and maintained in a list of **frequently** used terms, the **frequently** used

terms being excluded from the search by not using terms in the list for means for excluding **frequently** used terms from being considered for the search, the **frequently** used terms being contained in the index and maintained in a list of **frequently** used terms, the **frequently** used terms being excluded from the search by not using terms in the list for ...

...for each of the search terms wherein the term importance varies according to inverse document **frequency** of the search term.

28 An associative text search and retrieval system, according to claim... the ranks are calculated using a formula that varies according to the square of the **frequency** in each of the text documents of each of the search terms.

35 An associative...

...system, according to claim 34, wherein the formula also varies according to an inverse document **frequency** of each of the search terms.

36 An associative text search and retrieval system, according...the ranks are calculated using a formula that varies according to the square of the **frequency** in each of the text documents of each of the search terms.

42 A method...

...system, according to claim 41, wherein the formula also varies according to an inverse document **frequency** of each of the search terms.

43 A method of operating an associative text search...by not including noise terms in the index;
excluding from being considered for the search **frequently** used terms contained in the index and maintained in a list of **frequently** used terms, the **frequently** used terms being excluded from the search by not using terms in the list for...

...for each of the search terms wherein the term importance varies according to inverse document **frequency** of the search term.

51

17/5,K/13 (Item 9 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT

00189570
MULTIMEDIA SEARCH SYSTEM
SYSTEME DE RECHERCHE MULTISUPPORT
Patent Applicant/Assignee:
ENCYCLOPAEDIA BRITANNICA INC,
REED Michael,
BESTICK Greg,
GREENHALGH Carol,
BASTIN Norman J,
CARLTON Ron,
FRANK Stanley D,
GOOD Dale,
HOLZMAN Carl,
JENSEN Ann,

KESTER Harold,
MAATMAN Dave,
MUNEVAR Edwardo,
ROGERS Derryl,
MANUEL Leo W II,
LARSEN Arthur,
NEEDHAM Cristopher D,

Inventor(s):
REED Michael,
BESTICK Greg,
GREENHALGH Carol,
BASTIN Norman J,
CARLTON Ron,
FRANK Stanley D,
GOOD Dale,
HOLZMAN Carl,
JENSEN Ann,
KESTER Harold,
MAATMAN Dave,
MUNEVAR Edwardo,
ROGERS Derryl,
MANUEL Leo W II,
LARSEN Arthur,
NEEDHAM Cristopher D,

Patent and Priority Information (Country, Number, Date):
Patent: WO 9106916 A1 **19910516**
Application: WO 90US6098 19901026 (PCT/WO US9006098)
Priority Application: US 89917 19891026
Designated States:
(Protection type is "patent" unless otherwise stated - for applications prior to 2004)
AT AT AU BB BE BF BG BJ BR CA CF CG CH CH CM DE DE DK DK ES ES FI FR GA GB GB GR GR HU IT JP KP KR LK LU LU MC MG ML MR MW NL NL NO RO SD SE SE SN SU TD TG US
Main International Patent Class: G06F-015/40
Publication Language: English
Fulltext Availability:
Detailed Description
Claims
Fulltext Word Count: 31890

English Abstract
A database search system (10) that retrieves multimedia information in a flexible, user friendly system. The search system (10) uses a multimedia database consisting of text, picture, audio and animated data. That database is searched through multiple graphical and textual entry paths. Those entry paths include an idea search (30), a title finder search (40), a topic tree search (60), a picture explorer search (50), a history timeline search (90), a world atlas search (80), a researcher's assistant search (100), and a feature articles search (70).

French Abstract
Un systeme de recherche a base de donnees (10) extrait des informations multisupport dans un systeme flexible convivial. Le systeme de recherche (10) utilise une base de donnees multisupport consistant en des donnees textuelles, d'images, audio et animees. La recherche dans cette base de donnees s'effectue par de multiples chemins d'entree graphiques et textuels. Ces chemins d'entree comprennent une recherche d'idees (30), une recherche de titres (40), une recherche de themes (60), une recherche d'exploration d'image (50), une recherche de lignes de temps d'histoire (90), une recherche d'atlas du monde (80), une recherche d'assistance au chercheur (100) et une recherche d'articles caracteristiques (70).

Patent and Priority Information (Country, Number, Date):
Patent: ... **19910516**
Fulltext Availability:
Detailed Description
Claims
Publication Year: **1991**

Detailed Description
... ROM technology as a dynamically interactive way of presenting material contained in books, encyclopaedias, magazines, **catalogs** , etc. CD-ROMS offer a set of characteristics that are unique for this purpose. First...user wants to search for articles on "chickens", but does not know these articles are **categorized** under "poultry". then the search will ultimately fail. Alternatively, if this user is too narrow...

...relationships to the query word,, i.e. "broader term" or "related term".

Conventional systems also **frequently** fail to operate properly, or at all if, inaccurate or fragmented information is available. For...

...and complexity. For example, in the article , "Another Look at Automatic Text Retrieval Systems," by **Gerard Salton** (" **Salton** I") a blueprint is described for automatic indexing for a natural language database, Salton I...of Smart are incorporated herein by reference.

"Term-weighting Approaches in Automatic Text Retrieval" by **Gerard Salton** and Christopher Buckley ("Salton III') describes various text-indexing techniques for assigning weighted values to...

...precision. The article notes that recall-enhancing and precision enhancing measures usually include a term- **frequenc** ' factor
y
and an inverse document **frequency** factor (a form of term discrimination which usually consists of the product of the term **frequency** and the inverse of the document **frequency** and a normalization factor for documents). The Salton II article then describes various retrieval system...elements, to access structured thesaurus terms
for query words ranked by the best and most **frequently** used retrieved data elements.

It is still a further object of the invention to provide...consists of the researcher's as
sistant program. The researcher's assistant contains subject matter **categories** divided into topics and the topics are then divided into assignments. Each assignment is then...historical person@, etc.
The Researcher's Assistant entry path 100 contains a set of research **categories** and sub-topics. Each topic, in turn, includes three assignments which are **classified** accord ing to level of difficulty, The user may then choose one of those assignments...set-up research questions for users. Researcher's Assistant contains numerous articles

divided into several **categories** . Each particular **category** consists of topics containing three assignments. The assign ments are marked according to their level topic **categories** . A number of other
options are available on the display's prompt line including Notes...

...Path 25, Last Screen 26, Tools 106 and the Main menu 120.

If a research **category** is selected 722 from display 720, then a list of research topics appears 724. The...

...is interested in doing research on "spiders" he then selects "Living Things" from the research **categories** at step 720. The term "spider" would then be selected from the topic list at...

...are listed.
Should the user want to then select a different topic from the same **category** , the Back function 118 is activated again, returning to the topic menu 724. Clicking Back 118 once more, returns to the menu of research **categories** where the user can select a new **category** of topics.

Any time the user is looking at the introductory text for a topic...

...function 736
which will move them ahead 738 to the next topic in the same **category** without having to return to the topic menu 740. To change a research **category** , however, the Back function 118 must be employed as described above.

Once the user has...use, when the user clicks on the entry path icon, a display of the feature **categories** for Feature Articles are presented 750. The user then selects an article 752 by mov...of all word occur
rences, is then calculated, Pass 2 also calculates a series of **frequency** variable for all words in a BRU, such that the ratios of the words per...invention employs a hierarchically organized thesaurus
that relates terms to broader and narrower terms.
The **classification** of the thesaurus consists of terms which are arranged so as to be differentiated from...

...term. The system
indicates these types (called "facets") by facet headers enclosed in brackets.

The **classification** system also allows for relationships to other terms to be defined where those terms cannot...

...cross-reference descriptions are
provided to characterize these relationships.

A final aspect of the thesaurus **classification** is that it allows for compound terms to identify their place in the index.

For...

...term is followed by a hierarchical list of related terms derived from the above-noted **classification4** Term types are defined by the following set of descriptions.

NOT NOTATION (a running number...Plant movement (both types of stimulus-response behavior)
The compound form is indicated in the **classification** by the use of a 11+11 sign,, (i.e. GLACIATION :. GLACIATION + EROSION or EROSION...

...GLACIATION
BT EROSION
BT GLACIATION
Some examples of syntactic compounds from different parts of the **classification** .

CRYSTAL
CRYSTAL + DEFORMATION
SLIP (crystall.)
GLACIATION
GLACIATION + EROSION
GLACIAL QUARRYING
LOCOMOTION (animal behavior)
LOCOMOTION + PRIMATES...

...medium through which they travel, or the shape they possess, Most facet indicators in the **Classification** are of the following kinds.

Period
Place

Claim
... said textual entry paths comprises a researcher's assistant entry path which contains subject matter **categories** divided into topics and said topics are further divided into research assign ments wherein said...said
textual entry paths comprises a researcher's assistant entry path which contains subject matter **categories** divided into topics and said topics are further divided into research assignments wherein said research...

...article in said search system; and
a researcher's assistant means which has subject matter **categories** divided into topics which are further divided into research assignments ordered by level of difficulty...information retrieval system according to claim

80 wherein said weights are determined based upon the **frequency** of a given query term in a given unit of said information to be retrieved divided by the **frequency** of occurrences of said informa tion to be retrieved for all units of the information...

...92 wherein said weighting step involves assigning azatio weight for each term based upon the **frequency** of said term's appearance in a unit of said information to be retrieved divided by the **frequency** of said term's appearance in said total of said informa tion to be retrieved...such that terms in said database have pre-defined weights based
upon term type and **frequency** in said database.

```
File 347:JAPIO Nov 1976-2005/Apr(Updated 050801)
         (c) 2005 JPO & JAPIO
File 350:Derwent WPIX 1963-2005/UD,UM &UP=200553
         (c) 2005  Thomson Derwent
File 344:Chinese Patents Abs Aug 1985-2005/May
         (c) 2005 European Patent Office
File 371:French Patents 1961-2002/BOPI 200209
         (c) 2002 INPI.  All rts. reserv.

Set     Items   Description
S1       6239   CATALOG? OR TAXONOM? OR AUTOCLASSIF? OR AUTOCATEGOR? OR AU-
            TOCATALOG?
S2      97578   CLASSIFY? OR CLASSIFIE? ? OR CLASSIFICAT? OR CATEGORI? OR -
            CATEGORY?
S3          0   SALTON(1N)(GERALD OR GERARD)
S4          0   AU=SALTON G?
?
```

```
File 696:DIALOG Telecom. Newsletters 1995-2005/Aug 23
         (c) 2005 Dialog
File  15:ABI/Inform(R) 1971-2005/Aug 23
         (c) 2005 ProQuest Info&Learning
File  98:General Sci Abs/Full-Text 1984-2004/Dec
         (c) 2005 The HW Wilson Co.
File 112:UBM Industry News 1998-2004/Jan 27
         (c) 2004 United Business Media
File 141:Readers Guide 1983-2004/Dec
         (c) 2005 The HW Wilson Co
File 484:Periodical Abs Plustext 1986-2005/Aug W2
         (c) 2005 ProQuest
File 608:KR/T Bus.News. 1992-2005/Aug 24
         (c)2005 Knight Ridder/Tribune Bus News
File 813:PR Newswire 1987-1999/Apr 30
         (c) 1999 PR Newswire Association Inc
File 613:PR Newswire 1999-2005/Aug 24
         (c) 2005 PR Newswire Association Inc
File 635:Business Dateline(R) 1985-2005/Aug 23
         (c) 2005 ProQuest Info&Learning
File 810:Business Wire 1986-1999/Feb 28
         (c) 1999 Business Wire
File 610:Business Wire 1999-2005/Aug 24
         (c) 2005 Business Wire.
File 369:New Scientist 1994-2005/Jun W1
         (c) 2005 Reed Business Information Ltd.
File 370:Science 1996-1999/Jul W3
         (c) 1999 AAAS
File  20:Dialog Global Reporter 1997-2005/Aug 24
         (c) 2005 Dialog
File 624:McGraw-Hill Publications 1985-2005/Aug 23
         (c) 2005 McGraw-Hill Co. Inc
File 634:San Jose Mercury  Jun 1985-2005/Aug 23
         (c) 2005 San Jose Mercury News
File 647:CMP  Computer Fulltext 1988-2005/Aug W1
         (c) 2005 CMP Media, LLC
File 674:Computer News Fulltext 1989-2005/Aug W2
         (c) 2005 IDG Communications

Set     Items   Description
S1         11   GERALD(1N)SALTON
S2         10   AU='SALTON, G.'
S3         23   AU='SALTON, GERALD':AU='SALTON, GERARD,'
S4          0   AU='SALTON G'
S5         44   S1:S3
S6     653007   CATALOG? OR TAXONOM? OR AUTOCLASSIF? OR AUTOCATEGOR? OR AU-
              TOCATALOG?
S7    2043146   CLASSIFY? OR CLASSIFIE? ? OR CLASSIFICAT? OR CATEGORI? OR -
              CATEGORY?
S8         11   S5 AND S6:S7
S9          9   RD (unique items)
S10   1475709   FREQUEN?
S11         0   S4 AND S10
S12        31   GERARD(1N)SALTON
S13        24   S12 AND (S6:S7 OR FREQUEN?)
S14        32   S9 OR S13
S15         6   S14/2000:2005
S16        26   S14 NOT S15
S17        18   RD (unique items)
? t17/3,k/all
```

17/3,K/1 (Item 1 from file: 15)
DIALOG(R)File 15:ABI/Inform(R)

01838087 04-89078
Search engines: The 1999 Conference
Feldman, Susan
Information Today v16n6 PP: 1, 76+ Jun 1999
ISSN: 8755-6286 JRNL CODE: IFT
WORD COUNT: 3098

...TEXT: least I did.

Trends

Three themes emerged during these two intense days: visualization, metadata and **categorization** , and pursuit of the elusive user.

Visualization was the star of this show. As James...

...usually similarity between documents is calculated using some variation of the vector space model that **Gerald Salton** pioneered.

Many of the presenters at this conference were experimenting with visualizations to help the...
...Both InXight's and TextWise's tools invite interaction. You are presented with top-level **categories** . Clicking on topics lets you drill down to more specific subjects, or, at the end of the road, to a list of documents in that **category** . How will people use this novel approach to browse for and explore information?

Metadata and **categorization** was by far the most surprising trend to those of us from the library world...

...at this return to the past, chaired a panel that examined whether the need for **categorization** was valid. It seems that the automatic **categorization** in use today is an attempt to solve two problems: multiple meanings of terms, and...

...understand the contents of either a search or a large collection of Web pages. Automatic **categorization** , which occurs at the document processing stage, rather than at the searching stage, makes sense...

...what is in the collection, as well as in the search results. In contrast, manual **categorization** , a process that was questioned in the '50s and '60s, is a laborintensive activity that...
...is added to the NL index, it does not delay access to the information. This **categorization** also produces Northern Light's custom folders, a discovery and navigation tool I like a...

...is using people to review sites for appropriateness, and also to assign them to a **category** . Its target audience is new arrivals on the Web. It is aimed at family usage...

...entertain humans, but computers just keep going. For this reason, they are considering adding automatic **categorization** .

Dan Miller from Ask Jeeves described their human-centered process. Ask Jeeves tries to answer...Callan, of the University of Massachusetts, pointed out that it is difficult to create good **categories** because they

overlap. Clear distinctions are hard to define. They require labor and insert lag time in the process. A list of 30,000 **categories** is difficult to navigate. Full-text searching is an attractive alternative. Most of the Web...

...make documents easier to find, and to understand how people search than to return to **categorization** .
The value of this surprising return to an old library approach will not be resolved...

...most popular sites for that query, or they are creating directories (hence the interest in **categorization**) to help the user find the right ballpark so that he can browse productively. Sullivan...is, the definitions adapt as fields and terminology change. One of the major problems for **catalogers** has always been that change makes any static **classification** system out of date. In Euroferret, alerting profiles are adapted and changed by observing what...

...the user and re-weighs the terms. Similarly, they use agent technologies to create appropriate **categories** by using training sets of documents that have been **classified** manually. They find that new documents are tagged with about 95 percent accuracy by this...

...maps documents to users, users to users, ads and products to users, and concepts to **categories** . It makes sense: The processes are the same. Only the purposes differ.

Predictions for the...

17/3,K/2 (Item 2 from file: 15)

01557245 02-08234
Honoring the best of Information Science: The 1997 ASIS Annual Awards
Anonymous
American Society for Information Science. Bulletin v24n2 PP: 5-9 Dec 1997/Jan 1998
ISSN: 0095-4403 JRNL CODE: BAS
WORD COUNT: 3929

...TEXT: right, Ray Schwartz accepts Best SIG Publication honors on behalf of the members of SIG/ **Classification** Research.

Best Chapter Publication-of-the-Year The 1997 Best Chapter Publication-ofthe-Year Award...

...III). The SIG last won the award in 1993.

SIG Publication-of-the-Year SIG/ **Classification** Research (CR) is the winner of the 1997 SIG Publication-of-theYear Award, in recognition of Volume 7 of Advances in **Classification** Research, proceedings from the 7th ASIS SIG/CR **Classification** Research Workshop. Edited by Paul Solomon, the papers are representative of the quality and depth...ASIS president and professor of library and information studies at Rutgers University, has received the **Gerard Salton** Award for Excellence in Information Retrieval. The award, presented by the Special Interest Group in...

17/3,K/3 (Item 3 from file: 15)
DIALOG(R)File 15:ABI/Inform(R)

01236054 98-85449
Inside ASIS
Anonymous
American Society for Information Science. Bulletin v22n4 PP: 2-6 Apr/May 1996
ISSN: 0095-4403 JRNL CODE: BAS
WORD COUNT: 4036

...TEXT: concept maps, based on the IR literature, such as the Belkin and Croft (ARIST, 1987) **classification** of retrieval techniques, to provide an overarching structure for the system as well as a...the Department of Information Science, University of Pittsburgh, has been named president-elect of the **Classification** Society of North America, an interdisciplinary organization that promotes the scientific study of **classification** and clustering.

Clifford Lynch, current ASIS president and director of library automation, University of California...

...University of Massachusetts; Jan O. Pedersen and Marti A. Hearst, Xerox PARC; and the late **Gerard Salton** , Cornell University.
Charles T. Meadow, professor in the Faculty of Information Studies, University of Toronto...

17/3,K/4 (Item 4 from file: 15)
DIALOG(R)File 15:ABI/Inform(R)

01095033 97-44427
Individual, Inc.
Basch, Reva
Link-Up v12n5 PP: 6-7 Sep/Oct 1995
ISSN: 0739-988X JRNL CODE: LUP
WORD COUNT: 1782

...TEXT: extra charge. These special requests are also used to tweak and tune user profiles. If **categories** of information that were once on the periphery appear, by the pattern of special orders...

...HeadsUp customers "roll their own," choosing relevant topics from a detailed listing of 800-plus **categories** .

These range from the very broad--Automotive Industry Overview, US Banking Report, Front Page Healthcare...

...order, the Individual, Inc. search engine automatically re-adjusts the weight and ranking of each **category** within your interest profile.

The basic HeadsUp cost of $29.95 per month includes five...
...a subset of newspapers and wire services. As with HeadsUp, you define your own subject **categories** , then scan incoming news briefs once a day and order the stories you want to...a search engine called SMART (System for Manipulation and Retrieval of Text), developed by Dr. **Gerard Salton** at Cornell University. Unlike the traditional AND/OR Boolean logic on which pioneering information retrieval...

17/3,K/5 (Item 5 from file: 15)
DIALOG(R)File 15:ABI/Inform(R)

01049531 96-98924
Beyond Boole: The next logical step
Davis, Charles H
American Society for Information Science. Bulletin v21n5 PP: 17-20 Jun/Jul 1995
ISSN: 0095-4403 JRNL CODE: BAS
WORD COUNT: 2750

ABSTRACT: Many online public-access **catalogs** (OPAC) now provide keyword Boolean searching. People already comfortable with the "and", "or" and "not...
...TEXT: information specialists have become victims of their own success. For example, many online public-access **catalogs** (OPACs) now provide keyword Boolean searching. People already comfortable with the AND, OR and NOT...

...interface has been suggested by Michael Buckland and others that would permit "filtering" by such **categories** as language and date of publication. However, while reducing he overall size of the retrieval...

...Time), software entrepreneurs also have ignored the needs of people who use large databases and **catalogs** . Interestingly, researchers investigated the "impediments" to enhanced retrieval systems several years ago. Although vendors had...more powerful technique than Boolean logic. The method described has nothing to do with word **frequencies** in documents, nor does it involve the assignment of values by indexers. It empowers searchers...9, 381-384.

Doszkocs, Tamas E. (1983). "CITE NLM: "Natural Language Searching in an Online **Catalog** ," Information Technology and Libraries 2, no. 4, 364-380.

-- (1986). "Natural Language Processing in Information...
...Mulvihill, J. and E.H. Brenner (1968). "Ranking Boolean Output," American Documentation 19, 204-205.

Salton , Gerard and Michael McGill (1983). Introduction to Modern Information Retrieval. New York: McGraw Hill.

Salton , Gerard (1975). Dynamic Library and Information Processing. Englewood Cliffs, NJ: Prentice-Hall.

Sommar, H.G. and...

17/3,K/6 (Item 6 from file: 15)
DIALOG(R)File 15:ABI/Inform(R)

00880396 95-29788
Filtered information services
McCleary, Hunter
Online v18n4 PP: 33-42 Jul 1994
ISSN: 0146-5422 JRNL CODE: ONL
WORD COUNT: 4145

...TEXT: to any corporate department that deals with external information. Ellen Slaby of SandPoint said they **frequently** work with libraries, but the decision to use Hoover, however, is often made by multiple...

...to so much as look up a descriptor. Individual's SMART software was developed by **Gerard Salton** of Cornell University. Using various algorithms, it weighs the relevance of each word and assigns...

...built their businesses around being archival sources, although they are adding more current information and **frequent** update schedules.

Dialog does offer real-time newswire files such as Knight-Ridder, Reuters and...publications.
HOW CUSTOMIZABLE IS THE NEWS FEED? Some services limit your "search strategy" to preset **categories** . For example, a product from Individual, Inc. called HeadsUp lets you choose your profile from...

17/3,K/7 (Item 7 from file: 15)
DIALOG(R)File 15:ABI/Inform(R)

00880395 95-29787
Intelligent agents
Roesler, Marina; Hawkins, Donald T
Online v18n4 PP: 18-32 Jul 1994
ISSN: 0146-5422 JRNL CODE: ONL
WORD COUNT: 5974

...TEXT: by the user.

Software does not necessarily need to have all these qualities to be **classified** as an intelligent agent. On the other hand, it is probably reasonable to say that...PowerNews software to create a user agent that, based on the user's selection of **categories** , will daily download news clips on the computer, business, financial or medical industries to the...

...the subscriber chooses.

Individual uses the SMART text retrieval and filtering technology developed by Professor **Gerard Salton** at Cornell University to select material for subscribers' personal newspapers. Sources of information include general... available, such as batteries for his camera.
*Nancy is an active business woman who travels **frequently** . She contracts the services of an alerting agent to check flight times. If the flight...

17/3,K/8 (Item 8 from file: 15)
DIALOG(R)File 15:ABI/Inform(R)

00817335 94-66727
Automatic structuring and retrieval of large text files
Salton, Gerard ; Allan, James; Buckley, Chris
Communications of the ACM v37n2 PP: 97-108 Feb 1994
ISSN: 0001-0782 JRNL CODE: ACM
WORD COUNT: 6145

Salton, Gerard ...

...TEXT: the use of thesauruses tailored to particular subject areas, and of preconstructed knowledge structures that **classify** the main entities of interest in a given subject area, and specify the relationships that...

...12, 23].

A high-performance term weighting system assigns large weights to terms that occur **frequently** in particular documents, but rarely on the outside, because such terms are able to distinguish...

...A typical term weight of this type, known as a tf x idf weight (term **frequency** times inverse document **frequency**) may be defined as

(Equation (1) omitted)

were w sub ik represents the weight of term T sub k assigned to document D sub i , tf sub ik is the **frequency** of occurrence of 1 term T sub k in D sub i (a **frequency** of 0 is assumed for terms not assigned to D sub i), N is the...

...being retrieved. (Without length normalization, the longer documents with more assigned terms and higher term **frequencies** would generate higher document similarities, and exhibit higher retrieval potential, than the shorter items.) [18...1239.

25. Wittgenstein, L. Philosophical Investigations. Basil Blackwell and Mott Ltd., Oxford, England, 1953.

CR **Categories** and Subject Descriptors: H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing; H.3.2 [Information Storage and Retrieval]: Record **Classification** ; H.3.3 [information Storage and Retrieval]: Information Search and Retrieval; H.5.1 [Information...

...Words and Phrases: Content analysis, full-text searching, information retrieval, text structuring

ABOUT THE AUTHORS:

Gerard Salton has been a professor of computer science at Cornell University since 1965. He directs the...

17/3,K/9 (Item 9 from file: 15)

00759245 94-08637
Fortune visits 25 cool companies
Sherman, Stratford; Deutschman, Alan; Kupfer, Andrew; Serwer, Andrew E; et al
Fortune v128n7 PP: 56-90 Autumn 1993
ISSN: 0015-8259 JRNL CODE: FOR
WORD COUNT: 9722

...TEXT: would be writing software to sort the data. For help he turned to Cornell professor **Gerard Salton** , who had designed software called Smart to solve complex text-analysis problems. Salton agreed to...rights in six big cities, including New York and L.A., to hundreds of radio **frequencies** that had been used for communicating with fleets of vehicles such as cabs and repair...or PBXs. These built-to-order boxes are so complex that you

have to lug **catalogues** and manuals on sales calls, then go back to the office to write up the fall, is Conquer, a kind of intelligent **catalogue** for buyers of computer gear. Say you need to equip an office with desktop PCs...

17/3,K/10 (Item 10 from file: 15)
DIALOG(R)File 15:ABI/Inform(R)

00639731 92-54671
Happy Birthday to MELVYL (Part 2) - Ten Years Later: A Retrospective Prospectus
Brownrigg, Edwin
Information Technology & Libraries v11n3 PP: 272-277 Sep 1992
ISSN: 0730-9295 JRNL CODE: JLA
WORD COUNT: 3133

ABSTRACT: In an essay entitled Online **Catalogs** : Through a Glass Darkly (1983), what was foreseen by Brownrigg and Lynch, based on their...

...Internet, the MARC standard, and the Linked Systems Protocol, such that patrons of one online **catalog** could access other online **catalogs** remotely. In 1983, a **classification** of public access library automation systems was propounded in a partial hierarchy of difficulty based...
...TEXT: to revisit prognostications that Clifford Lynch and I made in 1983 in our article "Online **Catalogs** : Through a Glass Darkly"(1) In that essay we sought to make generalizations and offer...

...volume set--four copies of it. Later, while being shown the library's new online **catalog** , complete with personal computer front-end, I discovered that a citation to Experimental Researches in Electricity was not in the **catalog** . My host explained that no rare books or manuscripts were represented in the "public" **catalog** : If casual readers knew that the library had such materials, they would want to use...

...providing substantially enhanced service to the user, but also creating some daunting copyright-related problems.

CATALOGING FOR ACCESS

I remember vividly a University of California (UC) Library Council meeting in 1980...

...was an earnest debate over inclusion of subject access in the still-developing online union **catalog** , yet to be named MELVYL. The prevailing wisdom was that a union **catalog** was used for known-item searching and that subject access therefore would be a waste...

...access in the online system, the debate ended for the moment.

A study of online **catalog** use sponsored by the Council on Library Resources (CLR) finally put the issue to rest...

...and A&I databases as was once the practice of many libraries in their card **catalog** some fifty years ago. Those libraries created author, title, and subject entries for journal articles...

...role.

...perspective of the user: identification and location of information materials which is what the MELVYL **catalog** today provides.

As the CLR survey showed, the user wants more than to identify an...

...automated systems, including the MELVYL system.

AUTOMATING THE REFERENCE LIBRARIAN

In 1983 we propounded a **classification** of public access library automation systems in a partial hierarchy of difficulty based on what they would provide to the patron(13):

1. Places to look for an answer (online **catalog**).
2. Places to look for an answer with evaluation (extended online **catalogs** that integrate bibliographies and reviews).

3. Material containing an answer (extended online **catalog** with document delivery/electronic publishing).

4. Answers.

5. Answers to poorly posed questions.

We concluded...

...the fringes.(14)

REFERENCES AND NOTES

1. Edwin B. Brownrigg and Clifford A. Lynch, "Online **Catalogs** : Through a Glass Darkly," Information Technology and Libraries 2, no.1:104-15 (Mar. 1983).

2. Users Look at Online **Catalogs** : Results of a National Survey of Users and Non-Users of Online Public Access **Catalogs** , Final Report to the Council on Library Resources (Berkeley, Calif.: Division of Library Automation and...

...4. J. C. R. Licklider, Libraries of the Future (Cambridge, Mass.: MIT Press, 1965).

5. **Gerard Salton** , Dynamic Information and Library Processing (Englewood, N.J.: Prentice-Hall, Inc., 1975).

6. Theodore Hines...

...Time, Space, and Other Things (Avon Books, 1975), p.139.

8. Brownrigg and Lynch, "Online **Catalogs** ."

9. Allan D. Pratt, The Information of the Image (Norwood, N.J.: Ablex, 1982).

10...

...July 1977).
11. Ibid., p.57.

12. Ibid., p.60.

13. Brownrigg and Lynch, "Online **Catalogs** ."

14, W. David Penniman, "Libraries and the Future: Critical Research Needs," Lazerow Lecture, Simmons College...

17/3,K/11 (Item 11 from file: 15)
DIALOG(R)File 15:ABI/Inform(R)

00375598 87-34432
Historical Note: The Past Thirty Years in Information Retrieval
Salton, Gerard
Journal of the American Society for Information Science v38n5 PP: 375-380 Sep 1987
ISSN: 0002-8231 JRNL CODE: ASI

Salton, Gerard

...ABSTRACT: which combinations of keywords attached to text items are used to replace the more conventional **classification** schedules and subject heading **catalogs** . Between 1957 and 1959, H. P. Luhn became the first to propose that the computer...

17/3,K/12 (Item 12 from file: 15)
DIALOG(R)File 15:ABI/Inform(R)

00118611 80-12562
Automatic Term Class Construction Using Relevance-A Summary of Work in Automatic Pseudoclassification
Salton, G.
Information Processing & Management v16n1 PP: 1-15 1980
ISSN: 0306-4573 JRNL CODE: IPM

Salton, G.

...DESCRIPTORS: **Classification** ;

ABSTRACT: Both term **classifications** and thesauri have been used for many purposes in automatic information retrieval. A thesaurus can...

...with respect to search requests. Several experimental studies undertaken to test the construction of term **classifications** based on external relevance judgments indicate that improvements are obtainable in retrieval effectiveness. ...

17/3,K/13 (Item 13 from file: 15)
DIALOG(R)File 15:ABI/Inform(R)

00095544 79-10555
Suggestions for Library Network Design
Salton, G.
Journal of Library Automation v12n1 PP: 39-52 March 1979
ISSN: 0022-2240 JRNL CODE: JLA

Salton, G.

...ABSTRACT: run in one direction. However, the system does allow each library to maintain its own **cataloging** and content identification methods. From a technical point of view, a library network organization presents...

17/3,K/14 (Item 14 from file: 15)
DIALOG(R)File 15:ABI/Inform(R)

00026357 75-04739
A THEORY OF TERM IMPORTANCE IN AUTOMATIC TEXT ANALYSIS
SALTON, G. ; YU, C. T
JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE V26 N1 PP: 33-44
JAN-FEB 1975
ISSN: 0002-8231 JRNL CODE: ASI

SALTON, G ...

...ABSTRACT: IN CONTENT ANALYSIS TO SINGLE WORDS, JUXTAPOSED WORDS AND PHRASES, AND WORD GROUPS OR THESAURUS **CATEGORIES** . EXPERIMENTAL RESULTS ARE GIVEN SHOWING THE EFFECTIVENESS OF THE TECHNIQUE. GRAPHS. TABLES. REFERENCES.

17/3,K/15 (Item 1 from file: 98)
DIALOG(R)File 98:General Sci Abs/Full-Text

03026575 H.W. WILSON RECORD NUMBER: BGSI95026575
Performance of text retrieval systems.
Harman, Donna
Buckley, Chris; Callan, Jamie
Science (Science) v. 268 (June 9 '95) p. 1417-20
DOCUMENT TYPE: Feature Article
SPECIAL FEATURES: bibl ISSN: 0036-8075
LANGUAGE: English
COUNTRY OF PUBLICATION: United States

ABSTRACT: Readers comment on Marc Damashek's article "Gauging similarity with n-grams: Language-independent **categorization** of text," which appeared in the February 10 issue. Damashek asserted that his n-gram...

...3 tests and that it has not yet been proven adequate for general information retrieval. **Gerald Salton** of Cornell University makes similar arguments in a separate letter. Damashek responds.

17/3,K/16 (Item 1 from file: 484)
DIALOG(R)File 484:Periodical Abs Plustext

03307533 (USE FORMAT 7 OR 9 FOR FULLTEXT)
A conference on visualizing subject access
Dorman, David
Computers in Libraries (ICLB), v17 n5, p18-20, p.2
May 1997
ISSN: 1041-7915 JOURNAL CODE: ICLB

DOCUMENT TYPE: News
LANGUAGE: English RECORD TYPE: Fulltext; Abstract
WORD COUNT: 1320

TEXT:
... of data could be visualized. He displayed slides of threedimensional graphs, made by exporting such **catalog** data as call number and date of publication from various library **catalogs** into an Excel spreadsheet, manipulating the data, and then displaying resulting groupings in graphical format...

...late 1950s and early 1960s by people such as Ed Stiles, Lauren Doyle, and, subsequently, **Gerard Salton** , he stated. What is new today is the sophistication of computer processing. We finally have...
...bibliographic and full text-in each database's native mode. After seeing such powerful "virtual **catalogs** ," I could only agree with Doszkocs' conclusion that Z39.50 is too little and too...

...meaning.
Milstead characterized current Internet search engines as using statistical processing-measuring the occurrence and **frequency** of terms in a document-almost exclusively. Effective information retrieval will take a combination of...

...Richard Greenfield, a consultant at the Library of Congress. Before demonstrating LC's new experimental **catalog** -which Doszkocs describes as already the best OPAC in the world, and which can be...more effective technical services processing.
Greenfield also questioned the role of the MARC standard in " **cataloging** " Internet resources. The creation of MARC records is too labor intensive, and thus too expensive, for **cataloging** the millions of Web pages that appear every month. After raising these questions, he went...

...post-search processing of subject searches can be used to greatly enhance the usefulness of **catalog** data. Our opaque OPACs could definitely use some good VIBEs.
For More Information. . The richness...

17/3,K/17 (Item 2 from file: 484)
DIALOG(R)File 484:Periodical Abs Plustext

02037991 (USE FORMAT 7 OR 9 FOR FULLTEXT)
A revolutionary new product or a new marketing strategy?
McCleary, Hunter
Online (ONL), v18 n4, p33-42, p.9
Jul 1994
ISSN: 0146-5422 JOURNAL CODE: ONL
DOCUMENT TYPE: Feature
LANGUAGE: English RECORD TYPE: Fulltext; Abstract
WORD COUNT: 3690 LENGTH: Long (31+ col inches)

TEXT:
... to any corporate department that deals with external information. Ellen Slaby of SandPoint said they **frequently** work with libraries, but the decision to use Hoover, however, is often made by multiple...

...to so much as look up a descriptor. Individual's SMART software was developed by **Gerard Salton** of Cornell University. Using various

algorithms, it weighs the relevance of each word and assigns...

...built their businesses around being archival sources, although they are adding more current information and **frequent** update schedules.

Dialog does offer real-time newswire files such as Knight-Ridder, Reuters and...publications.

How customizable is the news feed? Some services limit your "search strategy" to preset **categories** . For example, a product from Individual, Inc. called HeadsUp lets you choose your profile from...

17/3,K/18 (Item 1 from file: 635)
DIALOG(R)File 635:Business Dateline(R)

0467393 94-20676
Individual's interactive newspaper hits stride
Porter, Patrick L
MASS HIGH TECH (Watertown, MA, US), V12 N3 s1 p2
PUBL DATE: 940124
WORD COUNT: 5,015
DATELINE: Cambridge, MA, US

TEXT:

...that by developing learning interface agents.

But first I would like to hear about Dr. **Gerard Salton** and how you ran into SMART and then maybe we could talk about how SMART...subscription base but it was producing incremental revenue for them. And as people were seeing **frequent** hits in a particular magazine or newsletter they would in some cases actually want to...

...editor will do is if the story editor comes in multiple installments throughout the day, **frequently** there is a flash headline at 10 o'clock and at 10:05 a paragraph...
...the highest ranked 50 or so articles and do a second pass on those. And **frequently** , like with the Bell Atlantic/TCI merger, that story might come to us from 60...look at the last five days of what we sent out to that subscriber because **frequently** sources get out of phase. Reuters might have covered it yesterday but the Chicago Tribune...

...which goes typically to a corporation or a department in a corporation and it is **frequently** fed via electronic mail into a local area network so an enterprise or a business...

...itself based on the articles that you rate relevant. I takes the concepts that appear **frequently** in the relevant articles and increases their weight and reduces the weight of the concepts that appear **frequently** in the non-relevant articles. That feedback helps us deal with what we call type...
?

no term frequency

```
File   9:Business & Industry(R) Jul/1994-2005/Aug 23
         (c) 2005  The Gale Group
File  13:BAMP 2005/Aug W2
         (c) 2005  The Gale Group
File  16:Gale Group PROMT(R) 1990-2005/Aug 24
         (c) 2005 The Gale Group
File  47:Gale Group Magazine DB(TM) 1959-2005/Aug 24
         (c) 2005 The Gale group
File  88:Gale Group Business A.R.T.S. 1976-2005/Aug 23
         (c) 2005 The Gale Group
File 148:Gale Group Trade & Industry DB 1976-2005/Aug 24
         (c)2005 The Gale Group
File 160:Gale Group PROMT(R) 1972-1989
         (c) 1999 The Gale Group
File 275:Gale Group Computer DB(TM) 1983-2005/Aug 24
         (c) 2005 The Gale Group
File 570:Gale Group MARS(R) 1984-2005/Aug 24
         (c) 2005 The Gale Group
File 621:Gale Group New Prod.Annou.(R) 1985-2005/Aug 24
         (c) 2005 The Gale Group
File 636:Gale Group Newsletter DB(TM) 1987-2005/Aug 24
         (c) 2005 The Gale Group
File 649:Gale Group Newswire ASAP(TM) 2005/Aug 12
         (c) 2005 The Gale Group
File 436:Humanities Abs Full Text 1984-2004/Dec
         (c) 2005 The HW Wilson Co

Set     Items   Description
S1     729955   CATALOG? OR TAXONOM? OR AUTOCLASSIF? OR AUTOCATEGOR? OR AU-
            TOCATALOG?
S2    3344055   CLASSIFY? OR CLASSIFIE? ? OR CLASSIFICAT? OR CATEGORI? OR -
            CATEGORY?
S3         89   SALTON(1N)(GERALD OR GERARD)
S4          0   AU='SALTON G'
S5         16   AU='SALTON, GERALD':AU='SALTON, GERARD'
S6         59   S3:S5 AND (S1:S2 OR FREQUEN?)
S7         40   RD (unique items)
S8          6   S7/2000:2005
S9         34   S7 NOT S8
```

? t9/3,k/all

9/3,K/1 (Item 1 from file: 13)
DIALOG(R)File 13:BAMP
(c) 2005 The Gale Group. All rts. reserv.

00546743 Supplier Number: 23882647 (USE FORMAT 7 OR 9 FOR FULLTEXT)
"Just the Answers, Please": Choosing a Web Search Service
(A comparison of Web search services is offered; discusses searches run as part of evaluation)
Article Author(s): Feldman, Susan
Searcher, v 5, n 5, p 44+
May 1997
DOCUMENT TYPE: Journal; Cross comparison study ISSN: 1070-4795 (United States)
LANGUAGE: English RECORD TYPE: Fulltext; Abstract
WORD COUNT: 4846

(USE FORMAT 7 OR 9 FOR FULLTEXT)

TEXT:
...the query.

As you work with Web search engines, be sure to read the FAQs (**Frequently** Asked Questions, Help, How to Search, etc.) and use the functions they describe. Watch out...
...of cars, with the longer query, including some articles on women and cars, and some **classified** ads for cars from Miami. Changing the query to "car reviews" located many more pertinent...

...other Web searchers to locate its results. Then it merges them, groups them into broad **categories** by type of site, and removes duplicates. As one would expect, it located the same...

...Apparently two words may outweigh one.

* Find technical papers on the SMART system written by **Gerard Salton** .

Salton was the father of relevance ranking information retrieval. His papers, as well as a wealth...

...Since Web search engines don't have an author field, I searched on SMART and **Gerard Salton** . Inference Find seemed a logical place to start looking, since the titles of the papers...

9/3,K/2 (Item 1 from file: 16)
DIALOG(R)File 16:Gale Group PROMT(R)
(c) 2005 The Gale Group. All rts. reserv.

01287398 Supplier Number: 41503889 (USE FORMAT 7 FOR FULLTEXT)
DATA RETRIEVAL UPHEAVAL
InformationWeek, p46
August 20, 1990
Language: English Record Type: Fulltext
Document Type: Magazine/Journal; Tabloid; General Trade
Word Count: 1369

... time the best hope for determining the relevance of documents to

text searchers. Unfortunately, word **frequency** counts and other techniques proved too simplistic. But this technique has made something of a...for each client."

The Smart system was developed on the basis of research done by **Gerard Salton** , a professor of computer science at Cornell University in Ithaca, N.Y., and perhaps the...

...leading developer of statistical text systems. Smart is based on the concept of inverse document **frequency** ; a document's relevance to one's search is determined by the number of times...

9/3,K/3 (Item 1 from file: 47)

05497095 SUPPLIER NUMBER: 57046532 (USE FORMAT 7 OR 9 FOR FULL TEXT)
Template Mining for Information Extraction from Digital Documents.
CHOWDHURY, GOBINDA G.
Library Trends, 48, 1, 182
Summer, 1999
ISSN: 0024-2594 LANGUAGE: English RECORD TYPE: Fulltext
WORD COUNT: 9355 LINE COUNT: 00783

... rules are available for the end-user. To solve this problem, they have proposed a **classification** of association rule types that provides a general framework for the design of association rule...

...of information are very difficult to identify. However, Croft further commented that extraction of simple **categories** of information is practical and can be an important part of a text-based information...D-Lib Magazine

(Chen et al., 1996) Hsinchun Chen, Chris Schuffels, and Rich Orwig, "Internet **Categorization** and Search: A Machine Learning Approach," Journal of Visual Communication and Image Representation, Special Issue...

...Proceedings of Hawaii International Conference on System Sciences '96 (Best Paper Award, Digital Documents Track).

(**Salton** , 1989) **Gerard Salton** , Automatic Text Processing, Addison-Wesley, 1989.

(Weiss et al., 1996)Ron Weiss, David Gifford et...

...information professionals over the years, the most prominent ones being the MARC formats, the AACR2 **catalog** formats, subject headings lists (such as the LCSH), and **classification** schemes such as LC, DDC, UDC, and so on. Each of these schemes is constructed...

...used for bibliographic access and control for decades, there remains the question of how to **catalog** and index materials available on the Internet using these schemes. This has given rise to...

...thought that electronic documents need to be self-indexed (as opposed to the assignment of **cataloging** and indexing tags and value added by **cataloging** and indexing agencies or library staff). However, it is obvious that, in order for the...lub.lu.se/cgi-bin/nmdc.pl): title, creator, subject (keywords, controlled vocabulary, and classification), **description** (abstract and content description), publisher, contributor (other than the creator), date, type (category **of** the resource), format (HTML, Postscript, etc.), identifier (URL, string or number used to identify the ...

...T.; Gaizauskas, R.; & Wilks, Y. (1996). Evaluation of an algorithm for the recognition and classification **of** proper names. In COLING '96 (The 16th International Conference on Computational Linguistics, August 5-9...

9/3,K/4 (Item 2 from file: 47)
DIALOG(R)File 47:Gale Group Magazine DB(TM)

05397152 SUPPLIER NUMBER: 54804041 (USE FORMAT 7 OR 9 FOR FULL TEXT)
Search Engines: The 1999 Conference.
Feldman, Susan
Information Today, 16, 6, 1
June, 1999
ISSN: 8755-6286 LANGUAGE: English RECORD TYPE: Fulltext
WORD COUNT: 3316 LINE COUNT: 00271

... least I did.
Trends
Three themes emerged during these two intense days: visualization, metadata and **categorization** , and pursuit of the elusive user.
Visualization was the star of this show. As James...

...usually similarity between documents is calculated using some variation of the vector space model that **Gerald Salton** pioneered.
Many of the presenters at this conference were experimenting with visualizations to help the...

...Both InXight's and TextWise's tools invite interaction, You are presented with top-level **categories** . Clicking on topics lets you drill down to more specific subjects, or, at the end of the road, to a list of documents in that **category** . How will people use this novel approach to browse for and explore information?
Metadata and **categorization** was by far the most surprising trend to those of us from the library world...

...at this return to the past, chaired a panel that examined whether the need for **categorization** was valid. It seems that the automatic **categorization** in use today is an attempt to solve two problems: multiple meanings of terms, and...

...understand the contents of either a search or a large collection of Web pages. Automatic **categorization** , which occurs at the document processing stage, rather than at the searching stage, makes sense...

...what is in the collection, as well as in the search results. In contrast, manual **categorization** , a process that was questioned in the '50s and '60s, is a labor-intensive activity...

...is added to the NL index, it does not delay access to the information. This **categorization** also produces Northern Light's custom folders, a discovery and navigation tool I like a...
...is using people to review sites for appropriateness, and also to assign them to a **category** . Its target audience is new arrivals on the Web. It is aimed at family usage...

...entertain humans, but computers just keep going. For this reason, they are considering adding automatic **categorization** .
Dan Miller from Ask Jeeves described their human-centered process.

Ask Jeeves tries to answer...Callan, of the University of Massachusetts, pointed out that it is difficult to create good **categories** because they overlap. Clear distinctions are hard to define. They require labor and insert lag time in the process. A list of 30,000 **categories** is difficult to navigate. Full-text searching is an attractive alternative. Most of the Web...

...make documents easier to find, and to understand how people search than to return to **categorization** .

The value of this surprising return to an old library approach will not be resolved...

...most popular sites for that query, or they are creating directories (hence the interest in **categorization**) to help the user find the right ballpark so that he can browse productively. Sullivan...is, the definitions adapt as fields and terminology change. One of the major problems for **catalogers** has always been that change makes any static **classification** system out of date. In Euroferret, alerting profiles are adapted and changed by observing what...

...the user and re-weighs the terms. Similarly, they use agent technologies to create appropriate **categories** by using training sets of documents that have been **classified** manually. They fi nd that new documents are tagged with about 95 percent accuracy by...

...maps documents to users, users to users, ads and products to users, and concepts to **categories** . It makes sense: The processes are the same. Only the purposes differ.

Predictions for the...

9/3,K/5 (Item 3 from file: 47)
DIALOG(R)File 47:Gale Group Magazine DB(TM)
(c) 2005 The Gale group. All rts. reserv.

04698681 SUPPLIER NUMBER: 18928465 (USE FORMAT 7 OR 9 FOR FULL TEXT)
Literature retrieval for interdisciplinary syntheses.(Navigating Among the Disciplines: The Library and Interdisciplinary Inquiry)
White, Howard D.
Library Trends, v44, n2, p239(26)
Fall, 1996
ISSN: 0024-2594 LANGUAGE: English RECORD TYPE: Fulltext; Abstract
WORD COUNT: 10858 LINE COUNT: 00949

... services do this for the articles and papers they cover. Library of Congress or Dewey **classification** codes, properly interpreted, do it for monographs and serials. Other sets of markers, such as...

...and pile up. Such co-occurrences link the disciplines. Crude measures of interdisciplinarity are simply **frequency** counts of these co- occurrences.

Classification codes do not occur in this way since they are disciplinary division - mutually exclusive by...

...of interdisciplinarity in individual authors, we might note whether any books they have published are **classified** outside their primary disciplinary fields. Thus, a contributor to this issue of Library Trends has published books **classified** in library and information science, his primary field, and in philosophy, a field in which he was trained. His name is a marker that links their LC **classification** codes.

Z BD

Patrick Wilson Patrick Wilson
This could be read as evidence either of...

...information science and sociology of knowledge (Wilson, 1977, 1983). Other authors associated with the Z **classification** who have published books **classified** in other fields include William S. Cooper (1978) in the P **classification** and **Gerard Salton** (1988) in the QA **classification** . Of course, to establish the extent to which authors are actually commingling fields, we must...

...of that work across a disciplinary divide. Porter and Chubin (1985) call these "citations outside **category** " (COCs). They distinguish two sorts:

1. breadth of citation BY a given article (or journal or research **category**); and 2. breadth of citation TO a given article (or journal or research **category**).

These may be designated as outgoing and incoming citations respectively. Assume that article XYZ is assigned to a subject **category** - e.g., economics. It may well cite other works. If so, one may ask, Are any outgoing citations made to works **classified** in some other discipline - across the border, so to speak? Similarly, one can ask whether...

...Article XYZ Article MNO
cites cites
Article ABC Article XYZ

By declaring some (operationally defined) **category** as central and then aggregating "citations outside **category** " across many writings, one can determine what fields a given literature draws upon and what... Depending on the database, they may be descriptors, identifiers, concept codes, LC subject headings, LC **classification** codes, journal titles, authors, names, and so on - a variety of bibliographic markers. By default they are displayed high to low in order of **frequency** of co-occurrence; they may also be requested in alphabetical order.

This interconvertibility of terms...

...to say nothing of end-users. That is to convert one kind of marker, LC **classification** codes, into another, their associated LC subject headings, in the LC MARC-Books database, which covers books **cataloged** by the Library of Congress since 1968.

The **classification** code chosen is GN 365.9, which stands for "Biological determinism. Sociobiology." Sociobiology is itself...
...we can see something of its components and also its ties (as perceived by subject **catalogers**) with fields beyond its usual range of connotation.(2)

First we select all documents posted to the **classification** code (CA). A space is necessary between ...RANK to display the LC subject headings assigned to this set in order of their **frequency** . The standard Dialog code for subject headings is DE (for "descriptors"), and we ask that ...

...to start with an LC subject heading (DE) and then to rank all the LC **classification** and Dewey codes (CA) that co-occur with it:

? SELECT SOCIOBIOLOGY/DE
S2 285 SOCIOBIOLOGY/DE ? RANK CA CONT

The ten most **frequently** occurring class codes follow. Note that LC and Dewey class codes are mixed in the...a retrieval of 295 articles after duplicates were removed. These were ranked by their subject **categories** (SC) using the "Continuous" option.

? RANK SC CONT
The result is a very clear display...

9/3,K/6 (Item 4 from file: 47)
DIALOG(R)File 47:Gale Group Magazine DB(TM)

04128688 SUPPLIER NUMBER: 16098666 (USE FORMAT 7 OR 9 FOR FULL TEXT)
Filtered information services; a revolutionary new product or a new marketing strategy? (services that electronically filter unwanted information) (includes two related articles)
McCleary, Hunter
Online, v18, n4, p33(9)
July, 1994
ISSN: 0146-5422 LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT; ABSTRACT
WORD COUNT: 4512 LINE COUNT: 00370

... to any corporate department that deals with external information. Ellen Slaby of SandPoint said they **frequently** work with libraries, but the decision to use Hoover, however, is often made by multiple...

...to so much as look up a descriptor. Individual's SMART software was developed by **Gerard Salton** of Cornell University. Using various algorithms, it weighs the relevance of each word and assigns current information and **frequent** update schedules.

Dialog does offer real-time newswire files such as Knight-Ridder, Reuters and...publications.

How customizable is the news feed? Some services limit your "search strategy" to preset **categories** . For example, a product from Individual, Inc. called HeadsUp lets you choose your profile from...

9/3,K/7 (Item 5 from file: 47)
DIALOG(R)File 47:Gale Group Magazine DB(TM)

04128687 SUPPLIER NUMBER: 16098240 (USE FORMAT 7 OR 9 FOR FULL TEXT)
Intelligent agents; software servants for an electronic information world (and more!). (autonomous and adaptive computer programs operating within software environments) (includes two related articles)
Roesler, Marina; Hawkins, Donald T.
Online, v18, n4, p18(11)
July, 1994
ISSN: 0146-5422 LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT; ABSTRACT
WORD COUNT: 6136 LINE COUNT: 00511

... by the user.

Software does not necessarily need to have all these qualities to be **classified** as an intelligent agent. On the other hand, it is probably reasonable to say that...PowerNews software to create a user agent that, based on the user's selection of **categories** , will daily download news clips on the computer, business, financial or medical industries to the...

...the subscriber chooses.

Individual uses the SMART text retrieval and filtering technology developed by Professor **Gerard Salton** at Cornell University to select material for subscribers' personal newspapers. Sources of information include general...available, such as batteries for his camera.

* Nancy is an active business woman who travels **frequently** . She contracts the services of an alerting agent to check flight times. If the flight...

9/3,K/8 (Item 6 from file: 47)
DIALOG(R)File 47:Gale Group Magazine DB(TM)

04081833 SUPPLIER NUMBER: 15454956 (USE FORMAT 7 OR 9 FOR FULL TEXT)
Automatic analysis, theme generation, and summarization of machine-readable texts.
Salton, Gerard ; Allan, James; Buckley, Chris; Singhai, Amit
Science, v264, n5164, p1421(6)
June 3, 1994
ISSN: 0036-8075 LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT; ABSTRACT
WORD COUNT: 5829 LINE COUNT: 00476

Salton, Gerard ...
... kind is the well-known equation [f.sub.t] X 1/[f.sub.c] term **frequency** times inverse collection **frequency**), which favors terms with a high **frequency** (ft) in particular documents but with a low **frequency** overall in the collection ([f.sub.c]). Such terms distinguish the documents in which they...

9/3,K/9 (Item 7 from file: 47)
DIALOG(R)File 47:Gale Group Magazine DB(TM)

03898950 SUPPLIER NUMBER: 13828062 (USE FORMAT 7 OR 9 FOR FULL TEXT)
Natural language comes of age. (West Publishing Co.'s WIN (Westlaw is Natural)) (includes related articles on answers to questions about WIN and on Dow Jones and Company Inc.'s plan to use Personal Librarian in all its databases)
Pritchard-Schoch, Teresa
Online, v17, n3, p33(9)
May, 1993
ISSN: 0146-5422 LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT; ABSTRACT
WORD COUNT: 5661 LINE COUNT: 00458

... retrieval methods use weighting methods. The most common are weights that are based on the **frequency** of a term in a single document, and its **frequency** in the entire collection. Probabilistic retrieval is based on the premise that the best overall...ranking is based on five factors that Dr. Koll has determined to be important:
* The **frequency** of each query word within a document is an indicator of relevance.
* A match on uses just the 100 most **frequent** words from each document....
Matching is performed by tallying word occurrences, combining scores for different...in the discipline.
Van Rijsbergen, C.J. Information Retrieval. 2d ed. Butterworth-Heinemann, 1979. 218pp.
Salton , Gerald and Michael J. McGill. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.400pp.
Allen, James...

9/3,K/10 (Item 8 from file: 47)
DIALOG(R)File 47:Gale Group Magazine DB(TM)

03813998 SUPPLIER NUMBER: 13539682 (USE FORMAT 7 OR 9 FOR FULL TEXT)

Research in technical communication: perspectives and thoughts on the process. (Special Issue: Research in Technical Communication)
Pinelli, Thomas E.; Barclay, Rebecca O.
Technical Communication, v39, n4, p526(7)
Nov, 1992
ISSN: 0049-3155 LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT; ABSTRACT
WORD COUNT: 4048 LINE COUNT: 00345

... failing to meet the standards of "scientific inquiry" in both of these areas.
The most **frequent** criticisms focus on the first component. Much of the early technical communication research was not...

...contents we can define?
4. Is there a lack of feedback between researchers and practitioners?
Gerald Salton 's response appeared in the July issue of JASIS under the title "A Note About...exist against the free exchange of knowledge with others outside of the organization. Restriction, security **classification** , and proprietary claims to knowledge, not free exchange and open access, characterize this reward system...
...the curriculums of many U.S. colleges and universities in recent years, these programs are **frequently** denigrated by the so-called "liberal arts."
With the full expectation of gaining acceptance on...that are used in different disciplines, technical communicators in academic programs could collaborate closely and **frequently** with researchers in these various disciplines. The net effect of this collaboration should be technical...

...results of research into practice. Technical communicators in academe could also collaborate more closely and **frequently** with the directors and managers of technical communication programs in government and industry. This cooperation...

...Journal of the American Society for Information Science 35, no. 2 (March 1984): 137.
6. **Gerald Salton** , "A Note about Information Science Research," The Journal of the American Society for Information Science...

9/3,K/11 (Item 9 from file: 47)
DIALOG(R)File 47:Gale Group Magazine DB(TM)

03634117 SUPPLIER NUMBER: 11295897 (USE FORMAT 7 OR 9 FOR FULL TEXT)
Global text matching for information retrieval.
Salton, Gerard ; Buckley, Chris
Science, v253, n5023, p1012(4)
August 30, 1991
CODEN: SCIEAS ISSN: 0036-8075 LANGUAGE: ENGLISH
RECORD TYPE: FULLTEXT
WORD COUNT: 2703 LINE COUNT: 00222

Salton, Gerard ...
... remainder of the collection. Thus, the term weighting system should assign low weights to high- **frequency** terms that occur in many documents of a collection and high weights to terms that...

...ik] to term [T.sub.k] in document [D.sub.i] in proportion to the **frequency** of occurrence of a term in [D.sub.i] and in inverse proportion to the...

...to a document, varies widely. The same is true of the values of the occurrence **frequencies** , [F.sub.ik], of terms [T.sub.k] in [D.sub.i]. To give each...

...term [T.sub.k] in document [D.sub.i], [f.sub.ik] is the occurrence **frequency** of [T.sub.k] in [D.sub.i], N represents the number of documents in...

...The weights of Eq. 1 range from 0 to 1 and include an enhanced term **frequency** factor (0.5 + 0.5[f.sub.ik]/max[f.sub.ip]) that varies only between 0.5 for terms with zero **frequency** and 1 for the most **frequent** terms. In addition, an inverse collection **frequency** factor, log(N/[n.sub.k]), is used, which is large for terms that occur...

9/3,K/12 (Item 10 from file: 47)
DIALOG(R)File 47:Gale Group Magazine DB(TM)

03634110 SUPPLIER NUMBER: 11295883 (USE FORMAT 7 OR 9 FOR FULL TEXT)
Developments in automatic text retrieval.
Salton, Gerard
Science, v253, n5023, p974(7)
August 30, 1991
CODEN: SCIEAS ISSN: 0036-8075 LANGUAGE: ENGLISH
RECORD TYPE: FULLTEXT
WORD COUNT: 7738 LINE COUNT: 00636

Salton, Gerard
... of, or, but, the, and so on) would then be used to delete the high-**frequency** function words that are insufficiently specific to represent document content. A suffix-stripping routine would...

...documents from the remainder of the collection. This suggests that the best terms will occur **frequently** in particular documents, but rarely on the outside. Two main components of the term weight must therefore be distinguished: the **frequency** of occurrence of a term [T.sub.k] in a document [D.sub.i], also known as the term **frequency** , [idf.sub.k], of [T.sub.k] in [D.sub.i] and the inverse document **frequency** , [idf.sub.k] of term [T.sub.k], which varies inversely with the number of...
...factor, known as the (tf x idf) weight (4, 18).

In addition to the term **frequency** and inverse document **frequency** , the length of each document, measured by the number of assigned terms, must also be...example a typical sequence such as "Alphabetic (adjective) characters (plural noun) occurring (gerund) most (quantifier) **frequently** (adverb) in (preposition) running (gerund) text (noun) account (noun or verb) for (preposition) 85 to...Process. Manage. 26, 111 (1990).

[20] K. Sparck-Jones, J. Doc. 28, 11 (1972); Keyword **Classification** for Information Retrieval (Butterworths, London, 1971).

[21] M. Dillon and A. S. Gray, J. A...

9/3,K/13 (Item 11 from file: 47)
DIALOG(R)File 47:Gale Group Magazine DB(TM)

03616141 SUPPLIER NUMBER: 11137486 (USE FORMAT 7 OR 9 FOR FULL TEXT)
An information system for corporate users: wide area information servers.
Kahle, Brewster; Medlar, Art
Online, v15, n5, p56(5)

Sept, 1991
CODEN: ONLID ISSN: 0146-5422 LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT
WORD COUNT: 3237 LINE COUNT: 00263

... and Mead Data Central's NEXIS for published text, one application can access all three **categories** of information. The user isn't required to become familiar with several entirely different systems...focus on a single professional field such as patent law or medical research.

REFERENCES

[1] **Salton , Gerald** and McGill, Micheal. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[2] DowQuest promotional literature...

9/3,K/14 (Item 12 from file: 47)

03483000 SUPPLIER NUMBER: 09042732 (USE FORMAT 7 OR 9 FOR FULL TEXT)
Linking library users: a culture change in librarianship; let patrons evaluate materials - in your automated catalogs **. A daring proposal from a leading educator.**
Koenig, Michael E.D.
American Libraries, v21, n9, p844(4)
Oct, 1990
CODEN: AMLRB ISSN: 0002-9769 LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT
WORD COUNT: 2682 LINE COUNT: 00218

Linking library users: a culture change in librarianship; let patrons evaluate materials - in your automated catalogs **. A daring proposal from a leading educator.**

TEXT:

Let patrons evaluate materials - in you automated **catalogs** . A daring proposal from a leading educator.

... most authoritative hypotheses about the probable relationships between cytokines and thyroid epithelial cell growth.

Researchers **frequently** glance at an article's list of references to assess its authoritativeness before deciding whether...

...paper, librarians certainly could not encourage patrons to scribble marginalia in library books or on **catalog** cards. That was defacing library property, in many cases a prosecutable misdemeanor. And we certainly did not encourage users to add their own cards to our **catalogs** .

There were good reasons then for discouraging USD in library collections. For one thing, it...

...commentary. For another, there was no practical way to delete USD from a holding or **catalog** card for patrons who preferred their source materials in pristine condition. Of course, a library...bibliographic and text retrieval systems to our electronic bulletin boards - in other words, mesh library **catalogs** and full text databases into hypertext systems in which USD comprises a major component of...

...have overlooked.

In short, we need to change tour mindset from that of creators of **catalogs** for others to use, and for which we are exclusive suppliers of all the data, to providers of a **catalog** users are free to supplement.

There are three reasons why we haven't implemented such...
...thought about recording USD, We have captured the insights of authors and publishers, indexers and **catalogers** -but never users.
Third, and by far the most disturbing, is our cultural blind spot...

...Logically, what is the difference between an opinion expressed in a book or in our **catalogs** ? First Amendment rights should apply equally in either place.
In making their historic compromise, librarians have cast the **catalog** in the role of neutral descriptive tool (though Sanford Berman reminds us of how many...

...slip in unobserved(11). That neutral role is just an artifact of the printed card **catalog** . There is no reason why the **catalog** cannot contain judgmental data too, so long as it is clearly labeled as such. It... limited. A student at Rosary College may well want to know whether a title is **frequently** used and often put on reserve at Yale University. USD applications will reach their greatest...

...to local systems. In fact, there is now substantial concern that the advent of cheap **cataloging** data on CD-ROM may cripple the goose that laid the golden eggs of library...
...Cyril. "Cranfield Tests on Index Language Devices," ASLIB Proceedings (June 1967:p. 173-194). (6.) **Salton** , **Gerald** and Chris Buckley. "Parallel Text Search Methods," communications of the ACM (February 1988: p. 202...

...DESCRIPTORS: Library **catalogs** --

9/3,K/15 (Item 13 from file: 47)
DIALOG(R)File 47:Gale Group Magazine DB(TM)
(c) 2005 The Gale group. All rts. reserv.

03459919 SUPPLIER NUMBER: 09275287 (USE FORMAT 7 OR 9 FOR FULL TEXT)
Customized information: "No, I don't want 'all the news that's fit to print.'" (includes related information on Dialog Alert service) (column)
Hawkins, Donald T.
Online, v14, n5, p117(4)
Sept, 1990
DOCUMENT TYPE: column ISSN: 0146-5422 LANGUAGE: ENGLISH
RECORD TYPE: FULLTEXT
WORD COUNT: 2358 LINE COUNT: 00200

... ventures show promise.
INDIVIDUAL, INC. INDIVIDUAL, Inc., of Cambridge, MA, uses technology developed by Professor **Gerard** **Salton** of Cornell University to produce a custom news product entitled First! which delivers customized information ...

...INDIVIDUAL's software uses a thesaurus to analyze text; concepts are determined by noting the **frequency** and placement of terms (those mentioned in headlines or the beginning of an article are...

9/3,K/16 (Item 14 from file: 47)
DIALOG(R)File 47:Gale Group Magazine DB(TM)
(c) 2005 The Gale group. All rts. reserv.

03303420 SUPPLIER NUMBER: 07852213 (USE FORMAT 7 OR 9 FOR FULL TEXT)

Questing for the "DAO": DowQuest and intelligent text retrieval.
Weyer, Stephen A.
Online, v13, n5, p39(10)
Sept, 1989
CODEN: ONLID ISSN: 0146-5422 LANGUAGE: ENGLISH RECORD TYPE:

FULLTEXT
WORD COUNT: 5442 LINE COUNT: 00432

... my query:
dow jones new text retrieval business information service
Since the database is updated **frequently** and contains articles for approximately the past 6-12 months, your results (especially by the... Search 1,2,3. This is known as "document relevance feedback," an approach pioneered by **Gerard Salton** at Cornell. DowQuest uses the words (the most significant and **frequent** 300 out of approximately 3,000 words) in the first three stories to arrive at...is used: after common noise/stop words are eliminated, DowQuest uses just the 100 most **frequent** words from each document. Each processor basically asks itself, "Which documents do I own that...
...words in other very "similar" documents. Giving feedback via examples and then doing associative, adaptive **classification** is central to connectionist or neural network models, a hot but hyped AI research area... system, the "understand" more of a story. For example, the CONSTRUE system from Carnegie Group **classifies** business stories for financial traders (e.g., "bonds," "acquisitions," "traders") in a fashion somewhat similar... Esther "The Index is Key" Release 1.0, 881 (Jan. 31,1988): pp. 1-8.
Salton , Gerard and Buckley, Chris. "Parallel Text Search Methods." Communications of the ACM Vol. 31, No. 2...

9/3,K/17 (Item 15 from file: 47)
DIALOG(R)File 47:Gale Group Magazine DB(TM)

03090342 SUPPLIER NUMBER: 06770289 (USE FORMAT 7 OR 9 FOR FULL TEXT)
Measuring the accuracy of diagnostic systems.
Swets, John A.
Science, v240, n4857, p1285(9)
June 3, 1988
CODEN: SCIEAS ISSN: 0036-8075 LANGUAGE: ENGLISH
RECORD TYPE: FULLTEXT
WORD COUNT: 7373 LINE COUNT: 00575

... way is available for general use. The preferred way quantifies accuracy independently of the relative **frequencies** of the events (conditions, objects) to be diagnosed ("disease" and "no disease" or "rain" and...

...negative" (cell c). Data from a test of a diagnostic system consist of the observed **frequencies** of those four possible outcomes.
However, if we consider proportions rather than raw **frequencies** of the four outcomes, then just two proportions contain all of the information about the...

...major ones and the basis for an appropriate accuracy measure.
A measure independent of event **frequencies** . Converting raw **frequencies** to proportions in the way just described creates one of two fundamental attributes of a...considers only the true-positive and false-positive proportions, an accuracy measure ignores the relative

frequencies , or prior probabilities, of positive and negative events--defined, respectively, as (a + c)/N and...

...assigned a particular system for detecting cracks in metal to be specific to the relative **frequencies** of cracked and sound specimens chosen for the test sample.
A measure independent of the...

...calculating the suitable measure. A measure of accuracy that is independent both of the relative **frequencies** of the two events and of the decision criterion that is adopted for a positive...

...either a rating of likelihood that a positive event occurred--for example, on a five- **category** scale ranging from "very likely" to "very unlikely"--or effectively a continuous quantity, for example...

...the analyst can convert to a rating. Then, in analysis, one considers different numbers of **categories** as representing a positive response (4). Most of the data reported in this article were...accomplished graphically but is usually performed by a computer program that accepts as inputs the **frequencies** of positive and negative diagnoses for each alternative event that are observed for various criteria...

...optimal weights and estimated a probability of malignancy.
Each mammogram was rated on a five- **category** scale of likelihood that the lesion was malignant; the **frequencies** of the various ratings, as pooled over the six observers, are shown in columns 2...

...Fig. 2. (The points span the graph well enough to avoid much extrapolation--a five- **category** rating scale, which yields four points within the graph, is usually adequte; other diagnostic fields often yield more data points, for instance, weather forecasting, where 13 rating **categories** are the norm, and aptitude testing or information retrieval, where the analyst can often derive...Figure 4 summarizes the results obtained with a computer-based system at Harvard University by **Gerard Salton** and Michael Lesk (on the right) and results obtained with a manual library system at...shows summary values for both types of study. About ten published studies exist in each **category** . Most of the field studies were reviewed in the context of signal detection theory and ROC analysis (37), and both **categories** were reviewed for the U.S. Congress Office of Technology Assessment (38). I have calculated...

...certainly for every item in the test sample whether it is positive or negative. Incorrectly **classifying** test items will probably depress measures of accuracy.
How are truly guilty and truly innocent parties to be determined for tests of the polygraph? Judicial outcomes and panel decisions may **categorize** erroneously, and even confessions can be false. Hence, one may resort to the analog study...for each query. In some instances, the degree of relevance was estimated on a four- **category** scale. Other studies have drawn queries directly from documents in the file, a procedure that...

9/3,K/18 (Item 1 from file: 88)

03439733 SUPPLIER NUMBER: 15071512
Automatic structuring and retrieval of large text files. (Technical)
Salton, Gerard ; Allan, James; Buckley, Chris

Communications of the ACM, v37, n2, p97(12)
Feb, 1994
DOCUMENT TYPE: Technical ISSN: 0001-0782 LANGUAGE: English
RECORD TYPE: Fulltext; Abstract
WORD COUNT: 5147 LINE COUNT: 00528

Salton, Gerard ...
... the use of thesauruses tailored to particular subject areas, and of preconstructed knowledge structures that **classify** the main entities of interest in a given subject area, and specify the relationships that...

...12, 23].
A high-performance term weighting system assigns large weights to terms that occur **frequently** in particular documents, but rarely on the outside, because such terms are able to distinguish...

...A typical term weight of this type, known as a tf x idf weight (term **frequency** times inverse document **frequency**) may be defined as
[Mathematical Expressions Omitted] where [w.sub.ik] represents the weight of term [T.sub.k] assigned to document [D.sub.i], [tf.sub.ik] is the **frequency** of occurrence of term [T.sub.k] in [D.sub.i] (a **frequency** of 0 is assumed for terms not assigned to [D.sub.i]), N is the...

...being retrieved. (Without length normalization, the longer documents with more assigned terms and higher term **frequencies** would generate higher document similarities, and exhibit higher retrieval potential, than the shorter items.) [18...

9/3,K/19 (Item 2 from file: 88)
DIALOG(R)File 88:Gale Group Business A.R.T.S.
(c) 2005 The Gale Group. All rts. reserv.

02079718 SUPPLIER NUMBER: 06905223
ACM press database and electronic products - new services for the Information Age.
Fox, Edward A.
Communications of the ACM, v31, n8, p948(4)
Aug, 1988
ISSN: 0001-0782 LANGUAGE: English RECORD TYPE: Fulltext
WORD COUNT: 2426 LINE COUNT: 00243

... a dictionary and a glossary of terms might be produced building upon the Computing Reviews **classification** scheme. Or, an information retrieval resercher might like a disc or tape with a set...SU, has provided editorial suggestions. Finally, special thanks go to J.C.R. Licklider and **Gerard Salton** for training and encouraging me to work in this field and in related research areas.

9/3,K/20 (Item 3 from file: 88)
DIALOG(R)File 88:Gale Group Business A.R.T.S.
(c) 2005 The Gale Group. All rts. reserv.

02068511 SUPPLIER NUMBER: 06242226
Parallel text search methods. (technical)
Salton, Gerald ; Buckley, Chris
Communications of the ACM, v31, n2, p202(14)
Feb, 1988
DOCUMENT TYPE: technical ISSN: 0001-0782 LANGUAGE: English

RECORD TYPE: Fulltext; Abstract
WORD COUNT: 8299 LINE COUNT: 00812

Salton, Gerald ...
GERAR D **SALTON** and CHRIS BUCKLEY
BOOLEAN TEXT RETRIEVAL METHODS
Since punched card days, text searches have been...only in special circumstances when the files are very small.
(2) In some situations a **classified** or clustered file organization is available where records covering the same subject matter are grouped...

...This is true of many library files where the documents are grouped according to library **classification** numbers. For clustered collections, fast search strategies are available that can concentrate the search to...

...directly on the importance of the term in the document text-for example, on the **frequency** of occurrence of the term in the text-and indirectly on the number of documents...

...documents to which they are attached from the remainder of the collection-are those occurring **frequently** in individual text items, but rarely in the remainder of the collection. Accordingly, a typical the **frequency** of occurrence of term j in document D.sub.i,f.sub.j is the...

...used, the weight of term j in document D.sub.i increases as the term **frequency** tf.sub.ij increases and as the collection **frequency** of the term b decreases.
A dynamic query improvement technique, known as relevance feedback, is...special-purpose devices designed to provide fast execution of particular operations that must be performed **frequently** , and prove particularly time consuming under normal circumstances. An obvious candidate is the list merging...Sigma].sub.j 1/log f.sub.j, where f.sub.j is the collection **frequency** of term j (the number of documents to which term j is assigned), and the...

...top 15 or 20 documents in decreasing order of the sum of the inverse collection **frequencies** of all matching query terms.
* A relevance feedback step is also available in which the...

...each new query term j will once again be weighted according to its inverse collection **frequency** 1/log f.sub.j. The original relevance feedback eq. (5), which increases or decreases...

...the terms assigned to distinct documents. This accounts for the use of the individual term **frequencies** tf.sub.ij of term j m in documents Di, as illustrated earlier in the...

...and have a better chance of being retrieved than documents with fewer terms.
Neither term **frequency** weights nor length normalization are available in the CM methodology. Nevertheless, it has been claimed...43].)
Consider now the problem of retrieval effectiveness. The CM methodology uses only inverse collection **frequency** weights attached to query terms. This must be compared with the use of alternative term...

...vector processing environments. Three types of term weighting components must be distinguished:
(1) a term **frequency** component that is based on the **frequency** of occurrence of a term in a given document or query,
(2) a collection **frequency** component that is based on the number of

documents in a collection to which a...

...weighting component, a complete term weight specification is expressible as a triple, covering the term **frequency** , inverse collection **frequency** , and vector normalization components, respectively. The computation of the document retrieval values is completely specified...

...for the ntc (document) x atn (query) weighting system. In this case a normalized term **frequency** times inverse collection **frequency** known as a (tf x idf) weighting scheme is used for document terms, and an enhanced term **frequency** times inverse document **frequency** system characterizes the query terms. The enhanced **frequency** component used for query terms in Figure 9 is designed to ensure that these weights...

...effectiveness of the different methods.

The output of Table III confirms that the inverse collection **frequency** weighting used in the CM document ranking system is in fact useful in retrieval. In...

...x btn) represents the CM retrieval order according to the sum of the inverse collection **frequencies** of matching ...terms, but the document term weights are changed from binary (i.e., unweighted) to term **frequency** (TF) and (TF x IDF) weights. The document term weights are not accessible to the...

...of Tables III-V demonstrate that, although the CM query weighting system in inverse collection **frequency** is beneficial compared to a totally unweighted retrieval method, much greater improvements can be produced...

...query, In the last two feedback runs of Table VII, ntc (i.e., normalized term **frequency** times inverse collection **frequency**) weights are used for query reformulation purposes. Moreover, a feedback method introduced by Ide [13...in certain SDI (selective dissemination of information) systems and core swapping is not required too **frequently** , and when discriminating attribute vectors are usable in each parallel processing unit, a fast parallel...

...1973, pp. 121-125.

7. Croft, W.B. A model of cluster searching based on **classification** . InfSyst. 5, 3 (1980) , 189-195.

8. Faloutsos, C., and Christodoulakis, S. Signature files: An... Comput. C-28, 6 (June 1979), 446-458.

39. Sparck Jones, K. Some thoughts on **classification** for retrieval./. Doc. 26, 2 (June 1970), 89-101.

40. Sparck Jones, K. A statistical...

...D.L. Applications of the Connection Machine. Computer 20, 1 (Jan. 1987), 85-97.

CR **Categories** and Subject Descriptors: C.1.2 [Processor Architectures]: Multiple Data Stream Architectures (Multiprocessors)-parallel processors...

...evaluation, text searching

Received 3/87; revised 11/87; accepted 10/87

Authors' Present Address: **Gerard Salton** and Chris Buckley, Department of Computer Science, Cornell University, Ithaca, NY 14853.

Permission to copy...

9/3,K/21 (Item 1 from file: 148)

DIALOG(R)File 148:Gale Group Trade & Industry DB

07576484 SUPPLIER NUMBER: 15875813 (USE FORMAT 7 OR 9 FOR FULL TEXT)
Searching natural language systems: searchers know thy engine. (includes related article on new natural language search engines)
Feldman, Susan E.
Searcher, v2, n8, p34(5)
Oct, 1994
ISSN: 1070-4795 LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT
WORD COUNT: 4390 LINE COUNT: 00349

... document. Documents which have the highest number of query words, as well as the greatest **frequency** of occurrence of each word, rate as the most relevant. [For background reading on relevance...

...problems. Word stems are isolated and matched against words with the same stem.

5. Very **frequent** terms may be ignored entirely. WAIS ignores any term which has more than 20,000 hits in the database.

6. Inverse term **frequency** : the comparison of how **frequently** a term appears in a document with how often it appears in the database as a whole. Within certain limits, terms which are rare in the entire database, but appear **frequently** in a particular document, make that document more relevant. Curiously, very rare terms appear to be less useful for retrieval, just as very **frequent** words are. Most algorithms aim at the middle range of term **frequencies** .

7. Relevance feedback. From the original list of hits, the searcher can choose the terms...avoid repeating words and use synonyms instead. Obviously, unless an automatic thesaurus were used, the **frequency** count for a well-written article might be lower than one for a run-of...

...of information. These resources already appear in digital form, but usually in an unstructured, un- **cataloged** , and unindexed manner. To put them in usable form manually would require funding and staff...

...in the new software: Matthew Koll (Personal Library Software), Elizabeth Liddy (Syracuse University and TextWise), **Gerard Salton** and James Allan (SMART and Cornell University), and Judy Feder (ConQuest). They gave us some...fields or coding (e.g., sf= or dt=product review). Since it is based on **frequency** of words, one occurrence will not weigh high in the algorithm. If you need to...

9/3,K/22 (Item 2 from file: 148)
DIALOG(R)File 148:Gale Group Trade & Industry DB

07488272 SUPPLIER NUMBER: 15625785 (USE FORMAT 7 OR 9 FOR FULL TEXT)
Chemical, Chase subscribe to electronic news system. (Chemical Banking Corp., Chase Manhattan Corp.) (Brief Article)
Marjanovic, Steven
American Banker, v159, n148, p11(1)
August 3, 1994
DOCUMENT TYPE: Brief Article ISSN: 0002-7561 LANGUAGE: ENGLISH
RECORD TYPE: FULLTEXT
WORD COUNT: 388 LINE COUNT: 00030

... Smart software - a system for manipulation and retrieval of text - developed by Cornell University professor **Gerard Salton** .

Some 36 banks and investment banks subscribe. Richard Vancil, vice president of marketing at Individual...

...too few stories. Mr. Salton's design gives different weights to words based on how **frequently** they appear.

"It's an intelligent alternative to key-word searches," Mr. Vancil said.

First...

9/3,K/23 (Item 3 from file: 148)
DIALOG(R)File 148:Gale Group Trade & Industry DB

06717288 SUPPLIER NUMBER: 14354908 (USE FORMAT 7 OR 9 FOR FULL TEXT)
Fortune visits 25 cool companies. (Company Profile)
Fortune, v128, n7, p56(15)
Autumn, 1993
DOCUMENT TYPE: Company Profile ISSN: 0015-8259 LANGUAGE: ENGLISH
RECORD TYPE: FULLTEXT; ABSTRACT
WORD COUNT: 10728 LINE COUNT: 00847

9/3,K/24 (Item 4 from file: 148)
DIALOG(R)File 148:Gale Group Trade & Industry DB

05562190 SUPPLIER NUMBER: 11691163 (USE FORMAT 7 OR 9 FOR FULL TEXT)
ASIS sponsors symposium on full-text retrieval. (American Society for Information Science)
Information Today, v8, n11, p13(3)
Dec, 1991
ISSN: 8755-6286 LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT
WORD COUNT: 2379 LINE COUNT: 00201

... associative retrieval systems which are now emerging, based on the work of Ed Stiles and **Gerald Salton** in the early 1960s, are now providing solutions to that problem, moving us away from...

...output, extended Boolean, relevance feedback, associative retrieval, natural language processing, text understanding and data extraction.

Gerald Salton , Cornell University, opened the session with a review of the SMART Project, which began 30...

...full-text retrieval.

Julian Yochum, Logicon, discussed another innovative methodology for text-retrieval, the Least **Frequent** Trigram Algorithm. The technique was introduced in a system that ran hourly searches against a...
...more than an hour, they knew that another technique was needed," he said. The least **frequent** trigram breaks a query into three letter groups and scans ...document. Waltz also described work at the Census Bureau using Memory-Based Reasoning to automatically **classify** documents. Experiments have shown an accuracy rate between 80-90 percent, comparable with human indexers...

9/3,K/25 (Item 5 from file: 148)
DIALOG(R)File 148:Gale Group Trade & Industry DB

05522208 SUPPLIER NUMBER: 11552067 (USE FORMAT 7 OR 9 FOR FULL TEXT)
Identifying barriers to effective subject access in library catalogs .
Lancaster, F.W.; Connell, Tschera Harkness; Bishop, Nancy; McCowan, Sherry
Library Resources & Technical Services, v35, n4, p377(15)
Oct, 1991
ISSN: 0024-2527 LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT
WORD COUNT: 7700 LINE COUNT: 00655

Identifying barriers to effective subject access in library catalogs .

TEXT:

Identifying Barriers to Effective Subject Access in Library **Catalogs**

The replacement of the card **catalog** by the online **catalog** brought with it a great resurgence of interest in the problems of subject access in general. This is hardly surprising in view of the fact that the online **catalog** promised to offer subject search capabilities that were substantially better than those offered by its predecessor.

Many studies on how to improve subject searching in online **catalogs** have already been performed. The approaches most **frequently** investigated can be grouped into five broad **categories** :

1. Those that rely on improved or more flexible approaches to the searching of elements...

...fourth group of studies looks at the effect of making additional searching aids available to **catalog** users. Bates proposes two such tools that could be used in existing **catalogs** based on Library of Congress Subject Headings (LCSH) (an end-user thesaurus - basically a vast...

...Lester found that such an authority file had relatively little effect on the ability of **catalog** users to match their subject terms with LCSH headings, while Van Pulis and Ludy found...

...to perform keyword searches in complete bibliographic records.[12]

Many keyword searches in large online **catalogs** would be successful in the sense that they would retrieve relevant items. But they would...

...last several years have added significantly to out store of knowledge on the behavior of **catalog** users and the performance of the subject **catalog** in libraries.

In general, however, almost all of the studies suffer from the fact that...
...crude or simplistic measures of searching success. This is a problem that has always bedeviled **catalog** -use studies (e.g., Lancaster[19],[20]). It is comparatively easy to evaluate a "known item" search in a library **catalog** : either a user finds the item or does not. A subject search cannot be evaluated...

...Instead, one needs a measure of the degree of success of a search.

While excellent **catalog** -use studies have been performed in the past (e.g., Lipetz,[21] and Tagliacozzo and...

...successful if the user is able to match subject terminology with the terminology of the **catalog** (examples) of this approach can be found in the work of Bates[24],[25]). Clearly the **catalog** user selects one or more items (and presumably borrows them) as the result of a...

...but the evaluation criterion is still very unsatisfactory.

The quality of subject access in library **catalogs** cannot be improved from the results of studies based on such imperfect criteria. A

Authority Control in an Online **Catalog** ," Journal of Academic Librarianship 12: 277-83 (1986). [13.] **Gerard Salton** and Michael J. McGill, Introduction to Modern Information Retrieval (New York: McGraw-Hill, 1983). [14.] Tamas E. Doszkocs, "CITES NLM: Natural-Language Searching in an Online **Catalog** ," Information Technology and Libraries 2:364-80 (1983). [15.] Gautum Biswas and others, "Knowledge-Assisted ...Haven: Yale University Library, 1970). [22.] R. Tagliacozzo and M. Kochen, "Information-Seeking Behavior of **Catalog** Users," Information Storage and Retrieved 6:363-81 (1970). [23.] Lester, Coincidence of User Vocabulary. [24.] Marcia J. Bates, "Factors Affecting Subject **Catalog** Search Success," Journal of the American Society for Information Science 28:161-69 (1977). [25...

...75 (1977). [26.] It is possible that not everyone will be willing to accept that **catalog** users seek the "best" materials. Nevertheless, it is the contention of the authors that even...

...found that title words do make a significant difference in the subject searching of library **catalogs** . [29.] Karen Markey Drabenstott and others, "Analysis of a Bibliographic Database Enhanced with a Library **Classification** ," Library Resources & Technical Services 34:179-98 (1990). [30.] John J. Knightly, "Traditional Information Gathering...

...The idea that some form of subject bibliography should substitute for subject access through the **catalog** of an academic library is far from new (see, for example, Elmer Michael Schloeder, "Selective Subject **Cataloging** : A Preliminary Analysis of a Possible Means of Reducing the Bulk of the **Catalog** in the University Library" [M.A. diss., University of Chicago, Graduate Library School, 1945], and Wesley Simonton, "Duplication of Subject Entries in the **Catalog** of a University Library and Bibliographies in English Literature," College & Research Libraries 11:215-21...

...whereby library users can add their evaluations of what they have read to online library **catalogs** .

F.W. Lancaster is Professor, and Nancy Bishop and Sherry McCowan were master's students...

DESCRIPTORS: Online **catalogs** --...

... **Cataloging** --

9/3,K/26 (Item 6 from file: 148)
DIALOG(R)File 148:Gale Group Trade & Industry DB

05416266 SUPPLIER NUMBER: 11068758 (USE FORMAT 7 OR 9 FOR FULL TEXT)
Crisis in cataloging **revisited: the year's work in subject analysis, 1990.**
Young, James Bradford
Library Resources & Technical Services, v35, n3, p265(18)
July, 1991
ISSN: 0024-2527 LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT
WORD COUNT: 10951 LINE COUNT: 00970

Crisis in cataloging **revisited: the year's work in subject analysis, 1990.**

TEXT:
Crisis in **Cataloging** Revisited: The Year's Work in Subject Analysis, 1990
... Young
The integration of mainstream American library traditions of subject analysis with modern indexing and **classification** theory and their adaptation to an online environment are bringing about a revolution in the practice of subject analysis. The research literature published in 1990 in the following **categories** is examined: subject **cataloging** **classification** , **classification** in online systems, subject access, indexing, the online environment, special materials, and special subjects. The literature gives evidence of a second crisis in **cataloging** , which will require a reconsideration of conceptual foundations.

For some time now, those who follow events in subject analysis have observed a parallel to the crisis in **cataloging** articulated by Osborn in 1941. [1] Yee has reviewed the period in which reorganization at the Library of Congress (LC), attempting to deal with perceived "crisis in **cataloging** ," set the stage for profound changes in **cataloging** theory and its application. [2] The absence of an event in subject **cataloging** parallel to the historical development of descriptive **cataloging** standards is intriguing. The portions of Cutter's Rules for a Printed Dictionary **Catalog** [3] that were concerned with description and access gave rise to successive expansions sponsored by...
...official successor. An understanding of certain important events leading to the development of Anglo-American **Cataloguing** Rules might aid an understanding of the absence of a similar development in subject **cataloging** . Osborn's 1941 article "The Crisis in **Cataloging** " dramatically galvanized widespread dissatisfaction into a call for action. In response, a penetrating analysis and distillation of the essence of the Anglo-American **cataloging** tradition was found in Lubetzky's **Cataloging** Rules and Principles. [4] LC's commitment in commissioning and supporting this work over many...

...role of thoughtful research in library management. International agreements, embodied in the International Conference on **Cataloging** Principles and the International Standards for Bibliographic Description, emerged as a result. Pondering the potential...

...computers. Wittman quantifies characteristics of subheadings used in award-winning indexes that were found less **frequently** in other indexas. Booth looks in detail at consistency in MEDLINE indexing. A known item...

...been indexed twice. Fifty-seven references were retrieved, comprising twenty-eight pairs of duplicates. Four **categories** of descriptors - major descriptors, minor descriptors, subheadings, and check tags - were compared for depth and...

...are discussed. Svenonius (C), in a review of the new edition of Eric Coates' Subject **Catalogues** : Headings and Structure, reflects on the continuing interest of Coates' work in chain indexing theory...explains the possibilities for use of the Preserved Context Indexing System (PRECIS) in the online **catalog** . Her presentation may help to make this important system better understood in this country. Although it incorporates modern concepts of indexing and **classification** , PRECIS was not developed primarily for online retrieval. She details those syntactical features of PRECIS rooted in analytico-synthetic **classification** that support postcoordination and permit it to function very effectively in an online environment. She...

...Britain, Hancock-Beaulieu evaluates the impact of an OPAC on subject searching behavior at the **catalog** and at the shelves and compares the matching approach of OPAC subject searching to the...
...Drabenstott and Vizine-Goetz detail the use of search trees for subject searching in online **catalogs** . Trees would control system response to determine appropriate subject searching approaches to user queries, based ...

...the user. Summarizing the vocabularies for online subject searching of bibliographic databases other than library **catalogs** , Piternick provides a lengthy synthesis of the dramatic impact online search techniques have had on...

...need to provide further automated links and leads. Broadbent experimented with providing both dictionary and **classified** access in the online **catalog** . Alphabetical and **classified** indexes were generated from subject headings and their linked LCC class numbers found in a...

...MARC bibliographic records to assess the feasibility of providing both a dictionary and a classed **catalog** from data in existing **catalog** records. An effective **classified** **catalog** was not found to be possible without further **classification** .

Geyser questions whether the end user is able to perform advanced subject searches using an OPAC. She discusses some advantages of verbal **classification** , such as the use of keyword searching, Boolean operators, manipulation of terms in different fields...Geyser also mentions end users' reactions and proposes criteria for evaluating OPACs.

An article entitled " **Classification** and Indexing Meeting" reports on conference papers from the Paris 1989 IFLA meeting. There were...

...efficient online use, are reviewed. Dale has compiled a bibliography of subject access in online **catalogs** . This is an overview with annotations of selected items from a rapidly growing literature divided...

...of true bilingual OPAC searching, in which one search retrieves headings in both languages, through **classification** , authority links, and even an automatic translation module. They cite the implications for multinational databases...

...Library Journal 115, no.4:115 (Feb. 15,
1990). Library of Congress. Office for Subject **Cataloging**
Policy. Free-floating Subdivisions:
An Alphabetical Index. 2d ed. Washington,
D.C.: Library of Congress, **Cataloging**
Distr. Service, 1990. Library of Congress. Subject **Cataloging**
Division,
Processing Department. Library of
Congress Subject Headings. 13th ed.
Washington, D.C.: Library of...

...Approach to Bibliographic
Processing." Online Review
14:3-12 (1990). McAllister-Harper, Desretta. "Dewey Decimal
Classification in the Online Environment:
A Study of Libraries in North
Carolina." **Cataloging** & **Classification**
Quarterly 11, no.1:45-58 (1990). McCarthy, Constance. "A Reference
Librarian's
View of the Online Subject **Catalog** ."
Cataloging & **Classification** Quarterly 10,
no.1/2:203-11 (1989). Mandel, Carol A., Lee Leighton, and Robert...

...Bell Ringing." In
Wursten, ed., In Celebration of Revised
780: Music in the Dewey Decimal **Classification**
Edition 20, p.39-52. Michalak, Thomas J. "An Experiment in
Enhancing **Catalog** Records at Carnegie-Mellon
University." Library Hi Tech 8,
no.3:33-42 (1990). Molholt...

...eds., Beyond the Book: Extending
MARC for Subject Access, p.157-69. Murdock, Paul R. " **Cataloging**
Catalysis:
Toward a New Chemistry of Conscience,
Communication and Conduct in the Online
Catalog ." **Cataloging** & **Classification**
Quarterly 10, no.1/2:65-80 (1989). Noreault, Terry, Dean Nita, and
Jenny Kriss...

...90-93 (1990). Rolland-Thomas, Paule, and Gerard Mercure.
"Subject Access in a Bilingual Online **Catalogue** ."
Cataloging & **Classification**
Quarterly 10, no.1/2:141-63 (1989). **Salton** , **Gerard** , Christopher
Buckley, and
Maria Smith. "On the Application of Syntactic
Methodologies in Automatic Text
Analysis...

...Terms." Information Processing & Management
26:543-48 (1990). Saye, Jerry D. "The Library of Congress
Classification
System in an Online Environment:
A Reaction." **Cataloging** & **Classification**
Quarterly 11, no.1:27-36 (1990). Slack, Frances, and Anthony J. Wood.
"Subject
Searching...
...no.6:41-49 (1990). Smiraglia, Richard P. "Subject Access to
Archival Materials Using LCSH." **Cataloging**
& **Classification** Quarterly 11, no.

...Catherine L. "Intellectual Level as
a Search Enhancement in the Online Environment:
Summation and Implications."
Cataloging & **Classification** Quarterly 11,
no.1:89-98 (1990). Williamson, Nancy J. "The Role of **Classification**
in Online Systems." **Cataloging** & **Classification**
Quarterly 10, no.1/2:95-104 (1989). Wilson, Patrick, and Nick
Robinson. "Form
Subdivisions...

...Wursten, Richard B., comp. (A). In Celebration
of Revised 780: Music in the Dewey
Decimal **Classification** Edition 20. MLA
Technical Report, no.19. Canton, Mass.:
Music Library Assn., 1990.
(B). "Introduction." In Wursten, ed.,
In Celebration of Revised 780: Music in the
Dewey Decimal **Classification** Edition 20,
p.1-27. Yee, Martha. "Subject Access to Moving
Image Materials in a...

DESCRIPTORS: Online **catalogs** --...

...Subject **cataloging** --...

... **Classification** , Library of Congress...

... **Classification** , Dewey decimal

9/3,K/27 (Item 7 from file: 148)
DIALOG(R)File 148:Gale Group Trade & Industry DB

03698265 SUPPLIER NUMBER: 06716388 (USE FORMAT 7 OR 9 FOR FULL TEXT)
Online data base industry timeline.
Meadow, Charles T.
Database, v11, n5, p23(9)
Oct, 1988
ISSN: 0162-4105 LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT
WORD COUNT: 1942 LINE COUNT: 00180

... M.I.T.
* System for the Mechanical Analysis and Retrieval of Text (SMART)
begun by **Gerard** **Salton** at Harvard University * Chemical Abstracts
Service's Chemical Titles, becomes first
regularly computer-produced publication...

...50 states 1969 *Henriette Avram at Library of Congress produces MARC
data
interchange standard for **catalog** data * ARPANet, first
packet-switched data communications network,
developed by Robert Taylor and Lawrence Roberts...

...community, first major online dial-up service * OCLC, under Frederick
Kilgour, initiates first shared library
cataloging system * Pandex, the first commercial database, brought
up on DIALOG for
limited access
* Roger Summit...Illinois by Martha Williams

*First mention of "end-user" in
an ERIC or LISA abstract: " **Frequency** And Impact Of Spelling Errors...," by Charles P. Bourne
*Apple II personal computer
marketed
1978...

9/3,K/28 (Item 1 from file: 275)
DIALOG(R)File 275:Gale Group Computer DB(TM)
(c) 2005 The Gale Group. All rts. reserv.

01693468 SUPPLIER NUMBER: 15567271 (USE FORMAT 7 OR 9 FOR FULL TEXT)
Letting computers choose your news. (Individual Inc's HeadsUp customized electronic news service)
Kador, John
MIDRANGE Systems, v7, n12, p39(1)
June 30, 1994
ISSN: 1041-8237 LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT; ABSTRACT
WORD COUNT: 1176 LINE COUNT: 00094

...ABSTRACT: create profiles that specify their particular interests, choosing items from a list of over 700 **categories** . HeadsUp utilizes a string of 486-based PCs and a software system called System for...
... recipients specified information. Subscribers create profiles by selecting topics from a list of over 700 **categories** . HeadsUp delivers information in two-sentence summaries of articles culled from hundreds of sources. Subscribers...

...interests.
Individual is the exclusive licensee for SMART, which was developed at Cornell University by **Gerald Salton** , a leading authority in artificial intelligence. SMART enables individuals' computers to sift through large volumes...

9/3,K/29 (Item 2 from file: 275)
DIALOG(R)File 275:Gale Group Computer DB(TM)
(c) 2005 The Gale Group. All rts. reserv.

01386529 SUPPLIER NUMBER: 09715931 (USE FORMAT 7 OR 9 FOR FULL TEXT)
The market for information markets. (the Knowledge Network that Reach Networks designed and implemented for partner Coopers & Lybrand provides broadcast-mode, data base-based information sharing for 40 percent of C&L's about-4,000 partners and managers)
RELease 1.0, v90, n11, p18(5)
Nov 30, 1990
ISSN: 1047-935X LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT
WORD COUNT: 3115 LINE COUNT: 00236

... analogies in favor of broadcasting, but suggests the analogy of an Official Airline Guide for **frequent** flyers -- choose what you want to watch.
C & L'S Knowledge Network
How do you...to produce shows -- which are up-to-the-week or up-to-the-day reports **classified** by practice area, written by outside writers for "clarity, wit and bubbliness." Rather than the...user hasn't yet seen this show.
Reach's technology includes a News Engine, which **classifies** news stories along the lines of NewsEDGE, First! or Third Eye (see Release 1.0

...
...it uses the same similarity-ranking approach based on the classic work of Cornell's **Gerard Salton** as the others (except Verity, which is based on topic hierarchies, and Reuters' **classification** scheme, which uses rules). However, says Stumm, "We don't see the value of going the last mile; we do just enough work to get useful **classifications** , not perfect ones [if there could be such a thing]. At Cornell they had one...

9/3,K/30 (Item 3 from file: 275)
DIALOG(R)File 275:Gale Group Computer DB(TM)
(c) 2005 The Gale Group. All rts. reserv.

01354664 SUPPLIER NUMBER: 08305728 (USE FORMAT 7 OR 9 FOR FULL TEXT)
Information refining. (software to summarize or perform semantic analysis on textual data) (includes related articles on language and meaning and how Reuters classifies **stories)**
RELease 1.0, v90, n3, p1(13)
March 16, 1990
ISSN: 1047-935X LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT
WORD COUNT: 5115 LINE COUNT: 00393

...**semantic analysis on textual data) (includes related articles on language and meaning and how Reuters** classifies **stories)**

TEXT:
...text if we know what we're looking for, but we need better ways to **categorize** texts and even summarize their meaning. We have covered many of them over the past...
... be; the more irregular, the more value can be added by refinement. Information can be **classified** , organized, filtered/selected and combed for relevant facts. Once refined, the output could be a...

...with practical problems; see the box across for a discussion of "meaning." The goal of **classification** or understanding of texts for our purposes is simply to represent a text as useful...

...into two basic but overlapping areas -- ones that deal with words and concepts, for text **classification** and filtering, and ones that try to derive summaries of the content (or "meaning"), generally...

...simple as a database or a list of index terms by which text chunks are **classified** , or as complex as a set of rules, a semi-hierarchical tree, a semantic net...

...object-oriented database or a set of situations and analogies. The model for filtering and **classification** systems is generally some set of relationships among lexical items; the model for the parsers...

...mixture of concepts, in fact).
The four sections following explore four approaches, three for text **classification** and the final one for summarization:
I - simple queries, no model;
II - simple queries, complex...

...Next, you can build more complex queries, which may weight words according to their relative **frequency** in the overall textbase and the target chunk of ...the weightings of words. They are all more or less based on the work of **Gerard Salton** at Cornell University.
Missing model

However, the complexity and power of the query mechanism has...

...a chronology or a table of contents or a course syllabus. A writer wanting to **categorize** source material according to a book outline would find a book outline useful; a student...

...Verity's Topic, a premier example of a system that helps users to build hierarchical **taxonomies** ("topics") and the related queries. It builds and displays relationships between topics and sub- and...variables ("acquired" followed by a company name) described below.

III -- Complex queries (with rules) for **classifying** texts by concept

We originally figured that sheer word statistics will always suffice to determine...

...passing mention. In tests at Reuters, for example, TIS got high recall (98 percent) for **categories** linked to specific names but lower precision (84 percent), because text items were flagged whether...

...Taylor,' he murmured," may not imply a story about the film industry. The 135 economic **categories** , by contrast, were tougher to recall (89 percent) than for precision because of the breadth...

...system described across provides such fine discrimination, assigning stories to a variety of carefully constructed **classifications** . By contrast, Desktop Data is using simpler string search (using customer-defined strings and information-vendor story **classifications** as available) on a pc to select news articles received in real-time over an...

...ignore.

IV -- Determining story lines: Complex models of individual text items

All the systems above **classify** text items, so that you can get appropriate pieces of text about any concept defined...builds a database about mergers and acquisitions, using stories from financial wires. Stories are first **classified** as about, not about or maybe about a takeover, in a four-stage process that...colleague George Krupka took the basic technology of SCISOR to build a second system for **classifying** Navy intelligence messages as part of a demonstration project for a DARPA conference, and were...

...a high-accuracy database generator in one person-month. (How accurate is, so to speak, **classified** .) The team has also started working with other GE units on possible commercial applications of...

...interfaces, for modification and reuse within Reuters. But Carnegie Group owns and licenses the underlying **categorization** and pattern-matching engine, Text **Categorization** Shell. If Reuters were to remarket TIS, it would owe royalties for TCS to Carnegie...

...DESCRIPTORS: **Taxonomy** ; ...

... **Classification** Systems

9/3,K/31 (Item 4 from file: 275)

01304028 SUPPLIER NUMBER: 07456938 (USE FORMAT 7 OR 9 FOR FULL TEXT)

Hypertext publishing: first steps.
RELease 1.0, v89, n6, p7(4)
July 13, 1989
ISSN: 1047-935X LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT; ABSTRACT
WORD COUNT: 2141 LINE COUNT: 00165

... using terms higher up in the hierarchy) and priorities. For example, "Macintosh" would belong to **categories** such as Apple and personal computers; "semiconductors" would include a long list of companies and...automatically with user feedback. The technique is similarity ranking, using techniques developed by Cornell's **Gerard Salton** , who sits on Individual's advisory board. These techniques are also the basis of Third...

...basis of further queries, appropriately weighted. That is, each word is assessed on its relative **frequency** , so that only words that appear much more **frequently** in the sample good story than they do in the entire database will be used...

9/3,K/32 (Item 5 from file: 275)
DIALOG(R)File 275:Gale Group Computer DB(TM)

01304026 SUPPLIER NUMBER: 07456826 (USE FORMAT 7 OR 9 FOR FULL TEXT)
The wonderful world of text. (overview of special issue on text software)
RELease 1.0, v89, n6, p1(3)
July 13, 1989
ISSN: 1047-935X LANGUAGE: ENGLISH RECORD TYPE: FULLTEXT; ABSTRACT
WORD COUNT: 12854 LINE COUNT: 01002

... another man's server
Note also that "client" software need not sit on "client" hardware. **Frequently** a "client" application will sit on the server, managing and manipulating the text files for...using terms higher up in the hierarchy) and priorities. For example, "Macintosh" would belong to **cat**egories such as Apple and personal computers; "semiconductors" would include a long list of companies and...automatically with user feedback. The technique is similarity ranking, using techniques developed by Cornell's **Gerard Salton** , who sits on Individual's advisory board. These techniques are also the basis of Third...

...basis of further queries, appropriately weighted. That is, each word is assessed on its relative **frequency** , so that only words that appear much more **frequently** in the sample good story than they do in the entire database will be used...reps "capture" new answers, and then "guide" them, or assign them by hand to appropriate **categories** for access by other support reps. (Lysis prefers to keep further details proprietary, and some ...

...to a variety of problems and by several paths. Ultimately, those that get used most **frequently** will bubble up to the top, just the way you might keep a crib sheet...

...answers generally gets into the system as text files that are chunked into "answers" and **classified** by the words in their headings. If the vendor already has documentation on line, so...T/Ware logic into the next release of Pulsar, a large-scale publishing system for **catalogue** operations used at Tandy and other major **catalogue** retailers. These customers design, copywrite, compose and assemble pages and proof on site,

and can...another man's server
Note also that "client" software need not sit on "client" hardware. **Frequently** a "client" application will sit on the server, managing and manipulating the text files for...

9/3,K/33 (Item 1 from file: 636)
DIALOG(R)File 636:Gale Group Newsletter DB(TM)
(c) 2005 The Gale Group. All rts. reserv.

01202657 Supplier Number: 41150893 (USE FORMAT 7 FOR FULLTEXT)
Electronic Publishing on Demand: Enhanced Fax and Other Media
Electronic Services Update, pN/A
Feb, 1990
Language: English Record Type: Fulltext
Document Type: Magazine/Journal; Trade
Word Count: 2103

... a technology search, the company became the exclusive licensee of indexing/search technology developed by **Gerard Salton** of Cornell, and it is now launching the first manifestation: "First!," a self-styled "Custom...

...a number of ways.) The Salton system also analyzes the placement of concepts, and their **frequency** , to gauge their weight and importance.
"That enables us to come out with a specific...

9/3,K/34 (Item 2 from file: 636)
DIALOG(R)File 636:Gale Group Newsletter DB(TM)
(c) 2005 The Gale Group. All rts. reserv.

01121094 Supplier Number: 40850744 (USE FORMAT 7 FOR FULLTEXT)
The User Interface: ASIS Midyear Meeting
Electronic Services Update, v2, n7, pN/A
July, 1989
Language: English Record Type: Fulltext
Document Type: Magazine/Journal; Trade
Word Count: 2342

... do we work alone. Experts are not expert in everything, and we have casual and **frequent** users."
Norman's seven stages of action require constraints to be minimized, determinations of what...the order of keyword occurrences).
His point was that different kinds of queries require different **category** sets wherein objects can be concepts, but not documents. Presearching sets and precoordinating series are...

...is connected to the document information store despite our front-end systems. Almost everything else -- **classification** , the partitioning of objects into sets, retrieval, and the interrelation between partitioning and a range...
...Documents are very physical." We look at physical characteristics and partitioning without yielding useful conceptual **categories** . "The purpose need not be document retrieval but educational or conceptual relationships," Miksa suggested, adding...

...University of California at Los Angeles made a single, important contribution when she stated that " **classification** is finally becoming floppy. Landmarks come when there are strong needs to create more precision

...

...and bring together concepts as distributed relatives represented in horizontal form rather than scattered in **classification** ." She concluded with this statement: "Hierarchical thinking is basic to thought."

Joseph A. Busch of...

...to be precise, but expert systems is a powerful capability."

Text Processing: Respectable and Booming

Gerard Salton , the famed professor of computer science and critic from Cornell who has spent his life...online world hasn't progressed much, enumerating a series of ills:

We have many online **catalogs** , but search mechanics are not easy, the errors are too many, there are too many...

...tirade against the industry, she enumerated the deficient tools of today's Online Public Access **Catalogs** (OPACs), particularly search problems, described as "systems problems, not user problems, and delays in giving...

...particularly for interactive dialog, online thesaurus aids for focusing, assistance in translating query terms into **cataloging** language, and the facilitation of open-ended and exploratory browsing." She also asked for offline...

?

```
File   6:NTIS 1964-2005/Aug W2
         (c) 2005 NTIS, Intl Cpyrght All Rights Res
File   2:INSPEC 1969-2005/Aug W2
         (c) 2005 Institution of Electrical Engineers
File   8:Ei Compendex(R) 1970-2005/Aug W2
         (c) 2005 Elsevier Eng.  Info. Inc.
File  57:Electronics & Communications Abstracts 1966-2005/Jul
         (c) 2005 CSA.
File  34:SciSearch(R) Cited Ref Sci 1990-2005/Aug W2
         (c) 2005 Inst for Sci Info
File  56:Computer and Information Systems Abstracts 1966-2005/Jul
         (c) 2005 CSA.
File  35:Dissertation Abs Online 1861-2005/Jul
         (c) 2005 ProQuest Info&Learning
File  60:ANTE: Abstracts in New Tech & Engineer 1966-2005/Jul
         (c) 2005 CSA.
File  65:Inside Conferences 1993-2005/Aug W3
         (c) 2005 BLDSC all rts. reserv.
File  94:JICST-EPlus 1985-2005/Jul W1
         (c)2005 Japan Science and Tech Corp(JST)
File  95:TEME-Technology & Management 1989-2005/Jul W3
         (c) 2005 FIZ TECHNIK
File  99:Wilson Appl. Sci & Tech Abs 1983-2005/Jul
         (c) 2005 The HW Wilson Co.
File 144:Pascal 1973-2005/Aug W2
         (c) 2005 INIST/CNRS
File 256:TecInfoSource 82-2005/Aug
         (c) 2005 Info.Sources Inc
File 266:FEDRIP 2005/Jun
         Comp & dist by NTIS, Intl Copyright All Rights Res
File 434:SciSearch(R) Cited Ref Sci 1974-1989/Dec
         (c) 1998 Inst for Sci Info
File 438:Library Lit. & Info. Science 1984-2005/Jul
         (c) 2005 The HW Wilson Co
File 583:Gale Group Globalbase(TM) 1986-2002/Dec 13
         (c) 2002 The Gale Group
File 439:Arts&Humanities Search(R) 1980-2005/Aug W2
         (c) 2005 Inst for Sci Info
File   1:ERIC 1966-2004/Jul 21
         (c) format only 2004 Dialog
File   7:Social SciSearch(R) 1972-2005/Aug W2
         (c) 2005 Inst for Sci Info
File 121:Brit.Education Index 1976-2005/Q4
         (c) 2005 British Education Index
File 142:Social Sciences Abstracts 1983-2005/Aug
         (c) 2005 The HW Wilson Co

Set     Items   Description
S1     359752   CATALOG? OR TAXONOM? OR AUTOCLASSIF? OR AUTOCATEGOR? OR AU-
             TOCATALOG?
S2    1610957   CLASSIFY? OR CLASSIFIE? ? OR CLASSIFICAT? OR CATEGORI? OR -
             CATEGORY?
S3         35   SALTON(1N)(GERALD OR GERARD)
S4        174   AU='SALTON G':AU='SALTON GG'
S5          3   AU='SALTON GJ'
S6        209   AU='SALTON, G':AU='SALTON, G.J.'
S7         95   AU='SALTON, GERALD':AU='SALTON, GERARD (ED. )'
S8        515   S3:S7
S9         37   S8 AND S1:S2
S10        32   S8 AND FREQUEN?
S11        69   S9:S10
```

S12 1 S11/2000:2005
S13 68 S11 NOT S12
S14 40 RD (unique items)

14/7/1 (Item 1 from file: 6)
DIALOG(R)File 6:NTIS

0440050 NTIS Accession Number: PB-231 539/8/XAB
A Theory of Indexing
(Technical rept)
Salton, G.
Cornell Univ., Ithaca, N.Y. Dept. of Computer Science.
Corp. Source Codes: 407072
Report No.: CU-CSD-74-203
Mar 74 102p
Journal Announcement: GRAI7412
Order this product from NTIS by: phone at 1-800-553-NTIS (U.S. customers); (703)605-6000 (other countries); fax at (703)321-8547; and email at orders@ntis.fedworld.gov. NTIS is located at 5285 Port Royal Road, Springfield, VA, 22161, USA.
NTIS Prices: PC A06/MF A01
Several automatic procedures are examined for the assignment of significance of values to the terms, or keywords, identifying the documents of a collection. Good and bad index terms are characterized by objective measures, leading to the conclusion that the best index terms are those with medium document **frequency** and skewed **frequency** distributions. A discrimination value model is introduced which makes it possible to construct effective indexing vocabularies by using phrase and thesaurus transformations to modify poor discriminations - those whose document **frequency** is too high, or too low - into better discriminators, and hence more useful index terms. Test results are included which illustrate the effectiveness of the theory. (Modified author abstract)

14/7/2 (Item 2 from file: 6)
DIALOG(R)File 6:NTIS

0419799 NTIS Accession Number: PB-226 063/6/XAB
Contribution to the Theory of Indexing
(Technical rept)
Salton, G. ; Yang, C. S. ; Yu, C. T.
Cornell Univ., Ithaca, N.Y. Dept. of Computer Science.
Corp. Source Codes: 407072
Report No.: CU-CSD-73-188
Nov 73 22p
Journal Announcement: GRAI7404
Order this product from NTIS by: phone at 1-800-553-NTIS (U.S. customers); (703)605-6000 (other countries); fax at (703)321-8547; and email at orders@ntis.fedworld.gov. NTIS is located at 5285 Port Royal Road, Springfield, VA, 22161, USA.
NTIS Prices: PC A02/MF A01
An attempt is made to characterize the usefulness of terms occurring in stored documents and user queries as a function of their **frequency** characteristics across the documents of a collection. It is found that the best terms are those having medium **frequency** in the collection and skewed **frequency** distributions. Correspondingly, terms exhibiting either very high or very low document **frequency** are not as useful. To improve the indexing vocabulary, it becomes necessary to group low **frequency** terms

into classes, and to break up high **frequency** terms by forming phrases. An indexing theory is described based on term **frequency** considerations, and a new phrase generation method is introduced. (Author)

14/7/3 (Item 3 from file: 6)
DIALOG(R)File 6:NTIS

0404044 NTIS Accession Number: PB-223 619/8/XAB
On the Specification of Term Values in Automatic Indexing
(Technical rept)
Salton, G. ; Yang, C. S.
Cornell Univ., Ithaca, N.Y. Dept. of Computer Science.
Corp. Source Codes: 407072
Report No.: CU-CSD-73-173
Jun 73 36p
Journal Announcement: GRAI7322
Order this product from NTIS by: phone at 1-800-553-NTIS (U.S. customers); (703)605-6000 (other countries); fax at (703)321-8547; and email at orders@ntis.fedworld.gov. NTIS is located at 5285 Port Royal Road, Springfield, VA, 22161, USA.
NTIS Prices: PC A03/MF A01
The existing practice in automatic indexing is reviewed, and it is shown that the standard theories for the specification of term values (or weights) are not adequate. New techniques are introduced for the assignment of weights to index terms, based on the characteristics of individual document collections. The effectiveness of some of the proposed methods is evaluated. (Author)

14/7/4 (Item 4 from file: 6)
DIALOG(R)File 6:NTIS

0214770 NTIS Accession Number: PB-188 957/XAB
Information Storage and Retrieval
(Scientific rept)
Salton, G.
Cornell Univ., Ithaca, N. Y. Dept. of Computer Science.
Report No.: ISR-16
Sep 69 496p
Journal Announcement: USGRDR7006
See also Report no. 15, PB-184 246.
Order this product from NTIS by: phone at 1-800-553-NTIS (U.S. customers); (703)605-6000 (other countries); fax at (703)321-8547; and email at orders@ntis.fedworld.gov. NTIS is located at 5285 Port Royal Road, Springfield, VA, 22161, USA.
NTIS Prices: PC A21/MF A01
Contract No.: NSF-750
The systems organization of the SMART programs is discussed as implemented for operation in a batch processing mode on the IBM 360/65. Covered in particular are the basic input and text analysis routines, the document clustering programs, the search routines and the feedback operations. Sample computer output is shown in each case to illustrate the operations. (Author)

14/7/5 (Item 5 from file: 6)
DIALOG(R)File 6:NTIS

0072698 NTIS Accession Number: PB-168 886/XAB
Progress in Automatic Information Retrieval
Salton, G.
Harvard Univ., Cambridge, Mass.
Corp. Source Codes: 163700
1965 15p
Document Type: Journal article
Journal Announcement: USGRDR6605
Pub. in IEEE Spectrum p90-103 Aug 1964.
Order this product from NTIS by: phone at 1-800-553-NTIS (U.S. customers); (703)605-6000 (other countries); fax at (703)321-8547; and email at orders@ntis.fedworld.gov. NTIS is located at 5285 Port Royal Road, Springfield, VA, 22161, USA.
NTIS Prices: PC A02/MF A01
The survey concentrates on word or text manipulating systems, in which items of information are represented by words in the matural language. Developments in computer organization, notably in time-sharing systems, are first reviewed briefly. This is followed by a description of current capabilities in automatic content analysis and content identification by statistical and structural methods, including word association techniques, term and document clustering, procedures using synonym dictionaries or hierarchical subject arrangements, and syntactic analysis. The design of a variety of automatic information systems is then reviewed-in particular, document retrieval systems, automatic technical centers, and questionanswering systems. An attempt is made to distinguish those system which appear to be technically and economically feasible from other which are likely to remain experimental for the foreseeable future. A prognosis is made of the type of automatic information system likely to become available within the next few years. (Author) Condensation of a paper given at the International Data Processing Conference and Business Exposition, Philadelphia, June 29-July 2, 1965)

14/7/6 (Item 6 from file: 6)
DIALOG(R)File 6:NTIS

0071358 NTIS Accession Number: PB-167 357/XAB
Automatic Information Processing in Western Europe
Salton, G.
Computation Lab., Harvard Univ., Cambridge, Mass.
Corp. Source Codes: 093650
1963 8p
Document Type: Journal article
Journal Announcement: USGRDR6401
Pub. in Science v144 n3619 p626-32 May 8 1964.
Order this product from NTIS by: phone at 1-800-553-NTIS (U.S. customers); (703)605-6000 (other countries); fax at (703)321-8547; and email at orders@ntis.fedworld.gov. NTIS is located at 5285 Port Royal Road, Springfield, VA, 22161, USA.
NTIS Prices: PC E03/MF A01
Current European work in automatic documentation and information processing is reviewed and evaluated.

14/7/7 (Item 1 from file: 2)
DIALOG(R)File 2:INSPEC

5759598 INSPEC Abstract Number: C9801-5260B-027
Title: Length normalization in degraded text collections
Author(s): Singhal, A.; **Salton, G.** ; Buckley, C.
Author Affiliation: Dept. of Comput. Sci., Cornell Univ., Ithaca, NY, USA
Conference Title: Proceedings. Fifth Annual Symposium on Document Analysis and Information Retrieval p.149-62
Publisher: Univ. Nevada, Las Vegas, NV, USA
Publication Date: 1996 Country of Publication: USA ix+336 pp.
Material Identity Number: XX96-00838
Conference Title: Proceedings of Fifth Annual Symposium on Document Analysis and Information Retrieval
Conference Sponsor: Inf. Sci. Res. Inst.; Howard R. Hughes Coll. Eng
Conference Date: 15-17 April 1996 Conference Location: Las Vegas, NV, USA
Availability: University of Nevada, Las Vegas, 4505 Maryland Parkway, Box 454021, Las Vegas, Nevada 89154-4021, USA
Language: English Document Type: Conference Paper (PA)
Treatment: Practical (P); Theoretical (T); Experimental (X)
Abstract: Optical character recognition (OCR) is the most commonly used technique to convert printed material into electronic form. Using OCR, large repositories of machine-readable text can be created in a short time. An information retrieval system can then be used to search through large information bases thus created. Many information retrieval systems use sophisticated term weighting functions to improve the effectiveness of a search. Term weighting schemes can be highly sensitive to the errors in the input text, introduced by the OCR process. This study examines the effects of the well known cosine normalization method in the presence of OCR errors, and proposes a new, more robust normalization method. Experiments show that the new scheme is less sensitive to OCR errors and facilitates the use of more diverse basic weighting schemes. This study also explains why the use of cosine normalization in presence of the inverse document **frequency** factor is not advisable in large document collections. (21 Refs)
Subfile: C

14/7/8 (Item 2 from file: 2)
DIALOG(R)File 2:INSPEC

5689688 INSPEC Abstract Number: C9710-7210-049
Title: Information astronomy on Internet: theory and practice
Author(s): Daranyi, S.
Journal: Tudomanyos es Muszaki Tajekoztatas vol.44, no.7-8 p.271-5
Publisher: OMIKK,
Publication Date: July-Aug. 1997 Country of Publication: Hungary
CODEN: TMTAAG ISSN: 0041-3917
SICI: 0041-3917(199707/08)44:7/8L.271:IAIT;1-T
Material Identity Number: T140-97007
Language: Hungarian Document Type: Journal Paper (JP)
Treatment: Practical (P)
Abstract: Information searching on the Internet is restricted by three facts: inadequate indexing, unclear search models and the misunderstanding called navigation. For true three- and four-dimensional navigation semantic universums should be prepared first which represent the knowledge of subject areas in space. The principles of this content mapping are to be found in the dynamic library idea of **Gerard Salton** . This recursive model can describe any continuously changing **classification** system in its evolution, and it is also suitable for handling electronic documents. Visual representation of information is also possible by substituting

cluster analysis in the original model by main component analysis. The result of the substitution is stable distributions of documents and keywords reminding one of astronomical constellations, leading to a still nonexistent information astronomy. (14 Refs)
Subfile: C

14/7/9 (Item 3 from file: 2)
DIALOG(R)File 2:INSPEC

03305567 INSPEC Abstract Number: C89014321
Title: Barriers to the use of research ideas in the design of real systems
Author(s): Dillon, M.
Conference Title: Influencing the System Designer: Online Public Access to Library Files: Third National Conference p.133-44
Editor(s): Dempsey, L.
Publisher: Elsevier Adv. Technol. Publications, Oxford, UK
Publication Date: 1988 Country of Publication: UK 163 pp.
ISBN: 0 946395 31 4
Conference Date: 12-15 Sept. 1987 Conference Location: Bath, UK
Language: English Document Type: Conference Paper (PA)
Treatment: General, Review (G)
Abstract: Discusses difficulties that arise in exploiting research ideas in online public access **catalogues** (OPACs), concentrating on the two major areas of design, system functionality and the user interface. In the area of functionality, the author focuses on automated feedback, an idea whose time came long ago, but which has made its way into few real systems. In order to deal with feedback, the author considers it in the context of **Gerard Salton** 's (1983) research on automatic text processing. With respect to the interface, the author discusses some reasons why interfaces evolve so slowly in interactive mainframe systems, in contrast to their rapid evolution in the microcomputer market. (11 Refs)
Subfile: C

14/7/10 (Item 4 from file: 2)
DIALOG(R)File 2:INSPEC

02126541 INSPEC Abstract Number: C83040051
Title: Automatic query formulations in information retrieval
Author(s): **Salton, G.** ; Buckley, C.; Fox, E.A.
Author Affiliation: Dept. of Computer Sci., Cornell Univ., Ithaca, NY, USA
Journal: Journal of the American Society for Information Science vol.34, no.4 p.262-80
Publication Date: July 1983 Country of Publication: USA
CODEN: AISJB6 ISSN: 0002-8231
U.S. Copyright Clearance Center Code: 0002-8231/83/040262-19$04.60
Language: English Document Type: Journal Paper (JP)
Treatment: Practical (P)
Abstract: Modern information retrieval systems are designed to supply relevant information in response to requests received from the user population. In most retrieval environments the search requests consist of keywords, or index terms, interrelated by appropriate Boolean operators. Since it is difficult for untrained users to generate effective Boolean search requests, trained search intermediaries are normally used to translate original statements of user need into useful Boolean search formulations. Methods are introduced in this study which reduce the role of

the search intermediaries by making it possible to generate Boolean search formulations completely automatically from natural language statements provided by the system patrons. **Frequency** considerations are used automatically to generate appropriate term combinations as well as Boolean connectives relating the terms. Methods are covered to produce automatic query formulations both in a standard Boolean logic system, as well as in an extended Boolean system in which the strict interpretation of the connectives is relaxed. Experimental results are supplied to evaluate the effectiveness of the automatic query formulation process, and methods are described for applying the automatic query formulation process in practice. (15 Refs)
Subfile: C

14/7/11 (Item 5 from file: 2)
DIALOG(R)File 2:INSPEC

01847206 INSPEC Abstract Number: C82019224
Title: A comparison of search term weighting: term relevance vs. inverse document frequency
Author(s): Wu, H.; **Salton, G.**
Author Affiliation: Cornell Univ., Ithaca, NY, USA
Journal: SIGIR Forum vol.16, no.1 p.30-9
Publication Date: Summer 1981 Country of Publication: USA
CODEN: FASRDV ISSN: 0163-5840
Conference Title: Proceedings of the Fourth International Conference on Information Storage and Retrieval
Conference Sponsor: ACM
Conference Date: 31 May-2 June 1981 Conference Location: Oakland, CA, USA
Language: English Document Type: Conference Paper (PA); Journal Paper (JP)
Treatment: Applications (A); Theoretical (T)
Abstract: The term relevance weighting method has been shown to produce optimal information retrieval queries under well-defined conditions. The parameters needed to generate the term relevance factors cannot unfortunately be estimated accurately in practice; furthermore, in realistic test situations, it appears difficult to obtain improved retrieval results using the term relevance weights over much simpler term weighting systems such as, for example, the inverse document **frequency** weights. It is shown that the inverse document **frequency** weights and the term relevance weights are closely related over a wide range of the **frequency** spectrum. Methods are introduced for estimating the term relevance weights, and experimental results are given comparing the inverse document **frequency** with the estimated term relevance weights. (15 Refs)
Subfile: C

14/7/12 (Item 6 from file: 2)
DIALOG(R)File 2:INSPEC

01832165 INSPEC Abstract Number: C82016205
Title: Term weighting in information retrieval using the term precision model
Author(s): Yu, C.T.; Lam, K.; **Salton, G.**
Author Affiliation: Univ. of Illinois, Chicago, IL, USA
Journal: Journal of the Association for Computing Machinery vol.29, no.1 p.152-70

Publication Date: Jan. 1982 Country of Publication: USA
CODEN: JACOAH ISSN: 0004-5411
Language: English Document Type: Journal Paper (JP)
Treatment: Applications (A); Theoretical (T)
Abstract: It is known that the use of weighted, as opposed to binary, content identifiers attached to the records of an information file improves the effectiveness of the retrieval operations. Under well-defined conditions the term precision offers the best possible term weighting system. A mathematical model is used in the present study to relate the term precision weights to the **frequency** of occurrence of the terms in a given document collection and to the number of relevant documents a user wishes to retrieve in response to a query. This provides for the assignment of user-dependent weights to the content identifiers and relates the term precision weights to other well-known weighting systems. (21 Refs)
Subfile: C

14/7/13 (Item 7 from file: 2)
DIALOG(R)File 2:INSPEC

01724569 INSPEC Abstract Number: C81027083
Title: The measurement of term importance in automatic indexing
Author(s): **Salton, G.** ; Wu, H.; Yu, C.T.
Author Affiliation: Dept. of Computer Sci., Cornell Univ., Ithaca, NY, USA
Journal: Journal of the American Society for Information Science vol.32, no.3 p.175-86
Publication Date: May 1981 Country of Publication: USA
CODEN: AISJB6 ISSN: 0002-8231
Language: English Document Type: Journal Paper (JP)
Treatment: Practical (P)
Abstract: The **frequency** characteristics of terms in the documents of a collection have been used as indicators of term importance for content analysis and indexing purposes. In particular, very rare or very **frequent** terms are normally believed to be less effective than medium- **frequency** terms. Recently automatic indexing theories have been devised that use not only the term **frequency** characteristics but also the relevance properties of the terms. The major term-weighting theories are first briefly reviewed. The term precision and term utility weights that are based on the occurrence characteristics of the terms in the relevant, as opposed to the nonrelevant, documents of a collection are then introduced. Methods are suggested for estimating the relevance properties of the terms based on their overall occurrence characteristics in the collection. Finally, experimental evaluation results are shown comparing the weighting systems using the term relevance properties with the more conventional **frequency** -based methodologies. (18 Refs)
Subfile: C

14/7/14 (Item 8 from file: 2)
DIALOG(R)File 2:INSPEC

01520664 INSPEC Abstract Number: C80019183
Title: Automatic term class construction using relevance-a summary of work in automatic pseudoclassification
Author(s): **Salton, G.**
Author Affiliation: Dept. of Computer Sci., Cornell Univ., Ithaca, NY, USA

Journal: Information Processing & Management vol.16, no.1 p.1-15
Publication Date: 1980 Country of Publication: UK
CODEN: IPMADK ISSN: 0306-4573
Language: English Document Type: Journal Paper (JP)
Treatment: Practical (P)
Abstract: Term **classifications** and thesauri can be used for many purposes in automatic information retrieval. Normally a thesaurus is generated manually by subject experts: alternatively, the associations between the terms can be obtained automatically by using the occurrence characteristics of the terms across the documents of a collection. A third possibility consists in taking into account user relevance assessments of certain documents with respect to certain queries in order to build term classes designed to retrieve the relevant documents and simultaneously to reject the nonrelevant documents. This last strategy, known as pseudoclassification, produces a user-dependent term **classification** . A number of pseudoclassification studies are summarized in the present report, and conclusions are reached concerning the effectiveness and feasibility of constructing term **classifications** based on human relevance assessments. (30 Refs)
Subfile: C

14/7/15 (Item 9 from file: 2)
DIALOG(R)File 2:INSPEC

01334389 INSPEC Abstract Number: C79013103
Title: Generation and search of clustered files
Author(s): **Salton, G.** ; Wong, A.
Author Affiliation: Cornell Univ., Ithaca, NY, USA
Journal: ACM Transactions on Database Systems vol.3, no.4 p.321-46
Publication Date: Dec. 1978 Country of Publication: USA
CODEN: ATDSD3 ISSN: 0362-5915
Language: English Document Type: Journal Paper (JP)
Treatment: Practical (P)
Abstract: A **classified** , or clustered file is one where related, or similar records are grouped into classes, or clusters of items in such a way that all items within a cluster are jointly retrievable. Clustered files are easily adapted to broad and narrow search strategies, and simple file updating methods are available. An inexpensive file clustering method applicable to large files is given together with appropriate file search methods. An abstract model is then introduced to predict the retrieval effectiveness of various search methods in a clustered file environment. Experimental evidence is included to test the versatility of the model and to demonstrate the role of various parameters in the cluster search process. (44 Refs)
Subfile: C

14/7/16 (Item 10 from file: 2)
DIALOG(R)File 2:INSPEC

01248443 INSPEC Abstract Number: C78025669
Title: Cataloging **software packages for automatic document processing**
Author(s): **Salton, G.**
Author Affiliation: Cornell Univ., Ithaca, NY, USA
Conference Title: Proceedings of the ASIS Annual Meeting 1976 (full papers available as microform only) p.80
Publisher: American Soc. Information Sci, Washington, DC, USA

Publication Date: 1976 Country of Publication: USA xi+176 pp.
ISBN: 0 87715 413 9
Conference Date: 4-9 Oct. 1976 Conference Location: San Francisco, CA, USA
Language: English Document Type: Conference Paper (PA)
Treatment: General, Review (G)
Abstract: While a good deal of attention has been devoted to the hardware aspects in information processing, relatively little is known about the software packages most appropriate for the new technology. A **classification** of important software routines useful in information retrieval and library processing is, however, as important as the standardisation of hardware products if the confusion stemming from the existence of the large multiplicity of almost similar, but actually incompatible, routines is to be alleviated. Eventually, 'off-the-shelf' information software could actually be made available to process the existing data collections. Various information processing tasks are described in this report together with the underlying data structure and the software packages necessary to perform the respective transformations.
Subfile: C

14/7/17 (Item 11 from file: 2)
DIALOG(R)File 2:INSPEC

01198494 INSPEC Abstract Number: C78015255
Title: Term relevance weights in online information retrieval
Author(s): **Salton, G.** ; Waldstein, R.K.
Author Affiliation: Dept. of Computer Sci., Cornell Univ., Ithaca, NY, USA
Journal: Information Processing & Management vol.14, no.1 p.29-35
Publication Date: 1978 Country of Publication: UK
CODEN: IPMADK ISSN: 0306-4573
Language: English Document Type: Journal Paper (JP)
Treatment: General, Review (G)
Abstract: Considerable evidence exists to show that the use of term relevance weights is beneficial in interactive information retrieval. Various term weighting systems are reviewed. An experiment is then described in which information retrieval users are asked to rank query terms in decreasing order of presumed importance prior to actual search and retrieval. The experimental design is examined, and various relevance ranking systems are evaluated, including fully automatic systems based on inverse document **frequency** parameters, human rankings performed by the user population, and combinations of the two. (12 Refs)
Subfile: C

14/7/18 (Item 12 from file: 2)
DIALOG(R)File 2:INSPEC

01086691 INSPEC Abstract Number: C77020381
Title: Effective information retrieval using term accuracy
Author(s): Yu, C.T.; **Salton, G.**
Author Affiliation: Univ. of Alberta, Edmonton, Alta., Canada
Journal: Communications of the ACM vol.20, no.3 p.135-42
Publication Date: March 1977 Country of Publication: USA
CODEN: CACMA2 ISSN: 0001-0782
Language: English Document Type: Journal Paper (JP)
Treatment: Theoretical (T)
Abstract: The performance of information retrieval systems can be

evaluated in a number of different ways. Much of the published evaluation work is based on measuring the retrieval performance of an average user query. Unfortunately, formal proofs are difficult to construct for the average case. In the present study, retrieval evaluation is based on optimizing the performance of a specific user query. The concept of query term accuracy is introduced as the probability of occurrence of a query term in the documents relevant to that query. By relating term accuracy to the **frequency** of occurrence of the term in the documents of a collection it is possible to give formal proofs of the effectiveness with respect to a given user query of a number of automatic indexing systems that have been used successfully in experimental situations. Among these are inverse document **frequency** weighting, thesaurus construction, and phase generation. (6 Refs)
Subfile: C

14/7/19 (Item 13 from file: 2)
DIALOG(R)File 2:INSPEC

00772429 INSPEC Abstract Number: C75014319
Title: A theory of term importance in automatic text analysis
Author(s): **Salton, G.** ; Yang, C.S.; Yu, C.T.
Author Affiliation: Cornell Univ. Ithaca, NY, USA
Journal: Journal of the American Society for Information Science vol.26, no.1 p.33-44
Publication Date: Jan.-Feb. 1975 Country of Publication: USA
CODEN: AISJB6 ISSN: 0002-8231
Language: English Document Type: Journal Paper (JP)
Treatment: Applications (A); Theoretical (T)
Abstract: Describes a new technique for automatic text analysis known as discrimination value analysis, ranks the text words in accordance with how well they are able to discriminate the documents of a collection from each other; that is, the value of a term depends on how much the average separation between individual documents changes when the given term is assigned for content identification. The best words are those which achieve the greatest separation. The discrimination value analysis is computationally simple, and it assigns a specific role in content analysis to single words, juxtaposed words and phrases and word groups of thesaurus **categories** . Experimental results are given showing the effectiveness of the technique. (16 Refs)
Subfile: C

14/7/20 (Item 14 from file: 2)
DIALOG(R)File 2:INSPEC

00758652 INSPEC Abstract Number: C75011625
Title: On the construction of effective vocabularies for information retrieval
Author(s): **Salton, G.** ; Clement, T.Yu.
Author Affiliation: Cornell Univ. Ithaca, NY, USA
Journal: SIGPLAN Notices vol.10, no.1 p.48-60
Publication Date: Jan. 1975 Country of Publication: USA
CODEN: SINODQ ISSN: 0362-1340
Conference Title: Proceedings of the ACM SIGPLAN-SIGIR interface meeting on programming languages-information retrieval
Conference Sponsor: Inst. Computer Sci. & Technol.; Nat. Bur. Standards
Conference Date: 4-6 Nov. 1973 Conference Location: Gaithersburg, MD,

USA
Language: English Document Type: Conference Paper (PA); Journal Paper (JP)
Treatment: Applications (A); Theoretical (T)
Abstract: Natural language query formulations exhibit advantages over artificial language statements since they permit the user to approach the retrieval environment without prior training and without using intermediaries. The usefulness of the terms in a natural language vocabulary is first characterized in terms of their **frequency** distribution over the documents of a collection. The construction of 'good' natural language vocabularies is then described, and methods are given for improving the vocabulary by transforming terms that operate poorly for retrieval purposes into better ones. (6 Refs)
Subfile: C

14/7/21 (Item 15 from file: 2)
DIALOG(R)File 2:INSPEC

00509422 INSPEC Abstract Number: C73011139
Title: Automatic processing of current affairs queries
Author(s): **Salton, G.**
Author Affiliation: Cornell Univ., Ithaca, NY, USA
Journal: Information Storage and Retrieval vol.9, no.3 p.165-80
Publication Date: March 1973 Country of Publication: UK
CODEN: IFSRAS ISSN: 0020-0271
Language: English Document Type: Journal Paper (JP)
Treatment: Experimental (X)
Abstract: The SMART system is used for the analysis, search and retrieval of news stories appearing in Time magazine. A comparison is made between the automatic text processing methods incorporated into the SMART system and a manual search using the **classified** index to Time. The results indicate that equivalent retrieval results are obtainable when both the manual and the automatic searches are carried out in a feedback mode. (14 Refs)
Subfile: C

14/7/22 (Item 16 from file: 2)
DIALOG(R)File 2:INSPEC

00170460 INSPEC Abstract Number: C70017654
Title: Automatic processing of foreign language documents
Author(s): **Salton, G.**
Author Affiliation: Cornell Univ., Ithaca, NY, USA
Journal: Journal of the American Society for Information Sciences vol.21, no.3 p.187-94
Publication Date: May 1970 Country of Publication: USA
CODEN: AISJB6 ISSN: 0002-8231
Language: English Document Type: Journal Paper (JP)
Abstract: Experiments conducted over the last few years with the SMART document retrieval system have shown that fully automatic text processing methods using relatively simple English language analysis tools are as effective for document indexing, **classification** , search, and retrieval as the more describes an extension of the SMART procedures to German language materials. A multilingual thesaurus is used for the analysis of documents and search requests, and tools are proved which make it possible to process English documents against German queries, and vice versa. The methods are

evaluated and it is shown that the effectiveness of the mixed language processing is approximately equivalent to that of the standard process operating within a single language only.
Subfile: C

14/7/23 (Item 17 from file: 2)
DIALOG(R)File 2:INSPEC
(c) 2005 Institution of Electrical Engineers. All rts. reserv.

00161180 INSPEC Abstract Number: C70015204
Title: Automatic text analysis
Author(s): **Salton, G.**
Author Affiliation: Cornell Univ., Ithaca, NY, USA
Journal: Science vol.168, no.3929 p.335-43
Publication Date: 17 April 1970 Country of Publication: USA
CODEN: SCIEAS ISSN: 0036-8075
Language: English Document Type: Journal Paper (JP)
Abstract: Automatic document indexing and **classification** methods are examined and their effectiveness is assessed. A review. (50 Refs)
Subfile: C

14/7/24 (Item 1 from file: 8)
DIALOG(R)File 8:Ei Compendex(R)
(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

01434415 E.I. Monthly No: EIM8308-055083
Title: AUTOMATIC INDEXING OF BUSINESS CORRESPONDENCE.
Author: **Salton, Gerard**
Corporate Source: Cornell Univ, Ithaca, NY, USA
Conference Title: Information Community: An Alliance for Progress, Proceedings of the 44th ASIS Annual Meeting.
Conference Location: Washington, DC, USA Conference Date: 19811025
Sponsor: ASIS, Washington, DC, USA
E.I. Conference No.: 02076
Source: Proceedings of the ASIS Annual Meeting 44th v 18. Publ for ASIS by Knowledge Industry Publ Inc, White Plains, NY, USA p 343
Publication Year: 1981
CODEN: PAISDQ ISSN: 0044-7870 ISBN: 0-914236-85-7
Language: English
Document Type: PA; (Conference Paper)
Journal Announcement: 8308

14/7/25 (Item 2 from file: 8)
DIALOG(R)File 8:Ei Compendex(R)
(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

01338532 E.I. Monthly No: EI8303018871 E.I. Yearly No: EI83046541
Title: ESTIMATION OF TERM RELEVANCE WEIGHTS USING RELEVANCE FEEDBACK.
Author: Wu, Harry; **Salton, Gerard**
Corporate Source: Cornell Univ, Ithaca, NY, USA
Source: Journal of Documentation v 37 n 4 Dec 1981 p 194-214
Publication Year: 1981
CODEN: JDOCAS ISSN: 0022-0418
Language: ENGLISH
Journal Announcement: 8303
Abstract: The term relevance weighting method has been shown to produce optimal information retrieval queries under well-defined conditions.

Unfortunately, the relevance weights cannot be determined in the absence of accurate knowledge of the occurrence **frequencies** of the terms in the relevant and non-relevant documents of a collection. This study presents a realistic method for estimating the term relevance weights from information derived in an interactive search environment where relevance assessments for previously retrieved items are used later to construct improved query statements. 23 refs.

14/7/26 (Item 3 from file: 8)
DIALOG(R)File 8:Ei Compendex(R)

01294912 E.I. Monthly No: EIM8305-029248
Title: TERM WEIGHTING MODEL BASED ON UTILITY THEORY.
Author: **Salton, Gerard** ; Wu, Harry
Corporate Source: Cornell Univ, Ithaca, NY, USA
Conference Title: Information Retrieval Research.
Conference Location: Cambridge, Engl Conference Date: 19800600
Sponsor: Br Comput Soc, London, Engl; ACM, New York, NY, USA
E.I. Conference No.: 01561
Source: Publ by Butterworths & Co Publ Ltd, London, Engl, and Boston, Mass, USA p 9-22
Publication Year: 1981
ISBN: 0-408-10775-8
Language: English
Document Type: PA; (Conference Paper)
Journal Announcement: 8305

14/7/27 (Item 4 from file: 8)
DIALOG(R)File 8:Ei Compendex(R)

00736968 E.I. Monthly No: EI7808056545 E.I. Yearly No: EI78020763
Title: CLUSTERED FILE GENERATION AND ITS APPLICATION TO COMPUTER SCIENCE TAXONOMIES .
Author: **Salton, Gerard** ; Bergmark, Donna
Corporate Source: Cornell Univ, Ithaca, NY
Source: Inf Process '77, Proc of IFIP (Int Fed for Inf Process) Congr, 6th, Toronto, Ont, Aug 8-12 1977 Publ by North-Holland Publ Co (IFIP Congr Ser, v 7), Amsterdam and New York, NY, 1977 p 441-445
Publication Year: 1977
Language: ENGLISH
Journal Announcement: 7808
Abstract: A clustered file organization is one where related, or similar records are grouped into classes, or clusters of items in such a way that all items within a cluster are jointly retrievable. Such a file organization is advantageous for interactive searching where tentative query formulations may be used and the records may be specified incompletely or approximately. An inexpensive file clustering method applicable to large files is given together with an appropriate file search method. The method is used to cluster a file of research articles in computer science based on citation similarities between the pacers; this leads to the identification of groups of active computer science research topics and of productive computer scientists. 8 refs.

14/7/28 (Item 5 from file: 8)
DIALOG(R)File 8:Ei Compendex(R)

00564128 E.I. Monthly No: EI7609060632 E.I. Yearly No: EI76036950
Title: AUTOMATIC INDEXING USING TERM DISCRIMINATION AND TERM PRECISION MEASUREMENTS.
Author: **Salton, G.** ; Wong, A.; Yu, C. T.
Corporate Source: Cornell Univ, Ithaca, NY
Source: Information Processing & Management v 12 n 1 1976 p 43-51
Publication Year: 1976
CODEN: IPMADK ISSN: 0306-4573
Language: ENGLISH
Journal Announcement: 7609
Abstract: The term discrimination model and the term precision system, two automatic indexing systems, are briefly described and experimental evidence is cited showing that a combination of both theories produces better retrieval performance than either one alone. The term discrimination model rates a given content indicator, or index term, in accordance with its usefulness as a discriminator among the documents of a collection, and offers in addition a reasonable physical interpretation for the indexing process. It is based largely on the occurrence characteristics and **frequency** distributions of the indexing units across the documents. Such a model does not account for the linguistic, or semantic aspects of the texts being processed, and may for this reason be criticized as simplistic and unrealistic. The term precision approach is designed to supply this missing dimension to some extent by utilizing customer opinions concerning the relevance or nonrelevance of certain documents to the queries submitted to the system. A precision weight attached to each query term is then used as a partial indication of the linguistic characterization of the terms. The procedures discussed must be more fully tested in operational environments where user relevance assessments obtained over a longer time period may provide reliable data for the term precision measurements. 9 refs.

14/7/29 (Item 1 from file: 34)
DIALOG(R)File 34:SciSearch(R) Cited Ref Sci

01185501 Genuine Article#: GC982 Number of References: 5
Title: GLOBAL TEXT MATCHING FOR INFORMATION-RETRIEVAL
Author(s): **SALTON G** ; BUCKLEY C
Corporate Source: CORNELL UNIV,DEPT COMP SCI/ITHACA//NY/14853
Journal: SCIENCE, 1991, V253, N5023, P1012-1015
Language: ENGLISH Document Type: ARTICLE
Abstract: An approach is outlined for the retrieval of natural language texts in response to available search requests and for the recognition of content similarities between text excerpts. The proposed retrieval process is based on flexible text matching procedures carried out in a number of different text environments and is applicable to large text collections covering unrestricted subject matter. For unrestricted text environments this system appears to outperform other currently available methods.

14/7/30 (Item 2 from file: 34)
DIALOG(R)File 34:SciSearch(R) Cited Ref Sci

01185494 Genuine Article#: GC982 Number of References: 97
Title: DEVELOPMENTS IN AUTOMATIC TEXT RETRIEVAL
Author(s): **SALTON G**

Corporate Source: CORNELL UNIV,DEPT COMP SCI/ITHACA//NY/14853
Journal: SCIENCE, 1991, V253, N5023, P974-980
Language: ENGLISH Document Type: ARTICLE
Abstract: Recent developments in the storage, retrieval, and manipulation of large text files are described. The text analysis problem is examined, and modern approaches leading to the identification and retrieval of selected text items in response to search requests are discussed.

14/7/31 (Item 1 from file: 99)
DIALOG(R)File 99:Wilson Appl. Sci & Tech Abs

1241392 H.W. WILSON RECORD NUMBER: BAST95037126
Performance of text retrieval systems
AUGMENTED TITLE: discussion and reply to the February 10, 1995 article, Gauging similarity with n-grams: language-independent **categorization** of text
Harman, Donna; Buckley, Chris; Callan, Jamie
Science v. 268 (June 9 '95) p. 1417-20
DOCUMENT TYPE: Feature Article ISSN: 0036-8075

ABSTRACT: Readers comment on Marc Damashek's article "Gauging similarity with n-grams: Language-independent **categorization** of text," which appeared in the 10 February issue. Damashek asserted that his n-gram information retrieval system, which searches document texts by looking for character strings instead of words or phrases, performed as well as some of the best existing retrieval systems at the third Text Retrieval Conference (TREC-3). Donna Harman and other members of the TREC Program Committee counter that Acquaintance was outperformed by most other entrants in the TREC-3 tests and that it has not yet been proven adequate for general information retrieval. **Gerald Salton** of Cornell University makes similar arguments in a separate letter. Damashek responds.

14/7/32 (Item 1 from file: 144)
DIALOG(R)File 144:Pascal

09002800 PASCAL No.: 90-0170981
On the use of spreading activation methods in automatic information retrieval
SALTON G ; BUCKLEY C
Cornell univ., dep. computer sci., Ithaca NY 14853, USA
International conference on research and development in information retrieval. 11 (Grenoble) 1988-06-13
1988 147-160
Publisher: Presses universitaires de Grenoble, Grenoble
Availability: CNRS-Y25163
No. of Refs.: 3 p.
Document Type: C (Conference Proceedings) ; A (Analytic)
Country of Publication: France
Language: English

14/7/33 (Item 2 from file: 144)
DIALOG(R)File 144:Pascal

01318299 PASCAL No.: 77-0073982
ON THE ROLE OF WORDS AND PHRASES IN AUTOMATIC TEXT ANALYSIS.
SALTON G ; WONG A
COMPUT. SCI. DEP.,CORNELL UNIV.
Journal: COMPUTERS AND HUMAN., 1976, 10 (2) 69-87
Availability: CNRS-14902
No. of Refs.: 14 REF.
Document Type: P (SERIAL) ; A (ANALYTIC)
Country of Publication: USA
Language: ENGLISH
ETUDE DU FONCTIONNEMENT DE L'ANALYSE AUTOMATIQUE DE TEXTES. PLUSIEURS THEORIES ET MODELES SONT EXAMINES QUI METTENT EN EVIDENCE LE PROCESSUS DE FORMATION DES PHRASES ET LA **FREQUENCE** STATISTIQUE DES MOTS.

14/7/34 (Item 3 from file: 144)
DIALOG(R)File 144:Pascal

00002659 PASCAL No.: 73-0002730
ON THE DEVELOPMENT OF INFORMATION SCIENCE
SALTON G
DEP. COMPUT. SCI., CORNELL UNIV., ITHACA,N.Y.
Journal: J. AMER. SOC. INFORM. SCI., 1973, 24 (3) 218-220
Availability: CNRS-6025
No. of Refs.: 5REF.
Document Type: P (SERIAL)
Country of Publication: USA
Language: ENGLISH
PAR UNE ETUDE COMPARATIVE DE DEUX PUBLICATIONS RECENTES EN SCIENCE DE L'INFORMATION, LE DEVELOPPEMENT ET L'ETAT ACTUEL DE CETTE SCIENCE SONT EXAMINES. CES PUBLICATIONS SONT L'INDEX CUMULATIF DE L'ARIST (ANNUAL REVIEW OF INFORMATION SCIENCE AND TECHNOLOGY) QUI COMPREND TOUTES LES REFERENCES CITEES DANS L'ARIST DU VOLUME 1 AU VOLUME 7, ET UNE BIBLIOGRAPHIE, LITERATUR ZU DEN INFORMATIONSWISSENSCHAFTEN, QUI EST UNE PARTIE D'UNE PUBLICATION EN TROIS VOLUMES, L'IBS (DAS INFORMATIONSBANKEN SYSTEM) ET QUI COMPREND DES ANALYSES EN ALLEMAND ET EN ANGLAIS. LES COMPARAISONS ONT ETE EFFECTUEES A PARTIR DES INDEX-AUTEURS. LES **FREQUENCES** DE CITATION DES PRINCIPAUX AUTEURS SONT NOTEES

14/7/35 (Item 1 from file: 1)
DIALOG(R)File 1:ERIC

00981310 ERIC NO.: ED414908 CLEARINGHOUSE NO.: IR056766
From **Classification** to "Knowledge Organization": Dorking Revisited or "Past is Prelude." FID Occasional Paper No. 14.
Gilchrist, Alan, Ed.;
CORP. SOURCE: International Federation for Information and Documentation, The Hague (Netherlands). (BBB30846)
191pp.
1997 (19970000)
NOTES: A collection of reprints to commemorate the forty year span between the Dorking Conference (First International Study Conference on **Classification** Research 1957) and the Sixth International Study Conference on **Classification** Research (London, UK) 1997.
REPORT NO.: FID-714
ISBN: 92-66-00-714-5
EDRS Price MF01 Plus Postage. PC Not Available from EDRS.

LANGUAGE: English
DOCUMENT TYPE: 20 (Collected works--General); 142 (Reports--Evaluative)
RECORD TYPE: ABSTRACT
COUNTRY OF PUBLICATION: Netherlands
JOURNAL ANNOUNCEMENT: RIEMAY1998

This set of papers offers insights into some of the major developments in the field of **classification** and knowledge organization, and highlights many of the fundamental changes in views and theories which have taken place during the last 40 years. This document begins with a series of reminiscences from former delegates of the first International Study Conference on **Classification** Research which took place in Dorking, United Kingdom in 1957, and continues with a collection of 15 papers by **classification** specialists: "The Need for a Faceted **Classification** as the Basis of all Methods of Information Retrieval"; " **Classification** in Information Retrieval: The Twenty Years Following Dorking" (E. J. Coates); "Structure and Function in Retrieval Languages" (B. C. Vickery); "Knowledge Representation: A Brief Review" (B. C. Vickery); "Natural Language Processing for Information Retrieval" (David D. Lewis and Karen Jones Sparck); "The Testing of Index Language Devices" (Cyril W. Cleverdon and J. Mills); "Indexing and Retrieval Performance: The Logical Evidence" (Dagobert Soergel); "Reflections on TREC" (Karen Sparck Jones); "On Information Science" (Carl Keren); "Brief Communication: A Note About Information Science Research" (**Gerard Salton**); "Unanswered Questions in the Design of Controlled Vocabularies" (Elaine Svenonius); "Needs for Research in Indexing" (Jessica L. Milstead); "Intelligent Text Processing, and Intelligence Tradecraft" (Michael L. Weiner and Elizabeth D. Liddy); "Advanced Searching: Tricks of the Trade" (Peggy Zorn, Mary Emanoil, Lucy Marshall, and Mary Panek); and "What do People Want from Information Retrieval?" (W. Bruce Croft). (SWC)

14/7/36 (Item 2 from file: 1)
DIALOG(R)File 1:ERIC

00264913 ERIC NO.: ED119707 CLEARINGHOUSE NO.: IR003175
Dynamic Information and Library Processing.
Salton, Gerard
523pp.
1975 (19750000)
AVAILABLE FROM: Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632 ($16.95)
Document Not Available from EDRS.
DOCUMENT TYPE: 10 (Book)
RECORD TYPE: ABSTRACT
JOURNAL ANNOUNCEMENT: RIEJUL1976

This book provides an introduction to automated information services: collection, analysis, **classification** , storage, retrieval, transmission, and dissemination. An introductory chapter is followed by an overview of mechanized processes for acquisitions, **cataloging** , and circulation. Automatic indexing and abstracting methods are covered, followed by a description of educational storage and retrieval systems. Library system analysis and evaluation are introduced in terms of theoretical models as well as practical applications. The final chapters of the book cover storage organization, automatic document and query **classification** , language processing, and dynamic information processing. At the end of each chapter is a bibliography. (CH)

14/7/37 (Item 3 from file: 1)
DIALOG(R)File 1:ERIC
(c) format only 2004 Dialog. All rts. reserv.

00211127 ERIC NO.: EJ112920 CLEARINGHOUSE NO.: IR501517
Automatic **Classification** : Directions of Recent Research
Moberg, Zandra
Drexel Library Quarterly, 10, 4, 90
1974 (19740000)
JOURNAL ANNOUNCEMENT: CIJJUL1975

14/7/38 (Item 4 from file: 1)
DIALOG(R)File 1:ERIC
(c) format only 2004 Dialog. All rts. reserv.

00138286 ERIC NO.: EJ068712 CLEARINGHOUSE NO.: LI502668
Comment on "An Evaluation of Query Expansion by the Addition of Clustered Terms for a Document Retrieval System"
Salton, G.
Information Storage and Retrieval, 8, 6, 349, Dec 72
1972 (19720000)
RECORD TYPE: ABSTRACT
JOURNAL ANNOUNCEMENT: CIJAPR1973

The author emphasized that one cannot conclude from the experiments reported upon that term clusters (or equivalently, keyword **classifications** or thesauruses) are not useful in retrieval. (2 references) (Author)

14/7/39 (Item 5 from file: 1)
DIALOG(R)File 1:ERIC
(c) format only 2004 Dialog. All rts. reserv.

00041702 ERIC NO.: ED027041 CLEARINGHOUSE NO.: LI001330
An Inquiry into Testing of Information Retrieval Systems. Comparative Systems Laboratory Final Technical Report, Part III: CSL Related Studies.
Zull, Carolyn Gifford, Ed.; And Others;
CORP. SOURCE: Case Western Reserve Univ., Cleveland, OH. Center for Documentation and Communication Research. (BBB00386)
151pp.
1968 (19680000)
NOTES: Related documents are ED 023 421, Part I of this final report, and LI 001 331, Part II.
SPONSORING AGENCY: Public Health Service (DHEW), Rockville, MD. (FGK74233)
CONTRACT/GRANT NO.: PHS-FR-00118
REPORT NO.: CSL-TR-FINAL-111
AVAILABLE FROM: Clearinghouse for Federal Scientific and Technical Information, Springfield, Va. 22151 (PB-180-952, MF $0.65, HC $3.00).
Document Not Available from EDRS.
RECORD TYPE: ABSTRACT
JOURNAL ANNOUNCEMENT: RIEJUL1969

This third volume of the Comparative Systems Laboratory (CSL) Final Technical Report is a collection of relatively independent studies performed on CSL materials. Covered in this document are studies on: (1) properties of files, including a study of the growth rate of a dictionary of index terms as influenced by number of documents in the file and a discussion of problems encountered in coding (**classifying**) different types of English words; (2) the nature of user questions which were searched in CSL, including a verbatim listing of the CSL questions and an

index to them as well as a **classification** of the questions according to various criteria; (3) the relations between the system answers (documents retrieved as answers) and the questions of the user, including a textual study of documents submitted as answers and a study which attempted to optimize searching of the CSL files on the basis of known relevant and nonrelevant answers; and (4) a comparison of the CSL study with similar experiments conducted by Cyril Cleverdon in the Cranfield II project and **Gerard Salton** at Cornell University. A list of CSL technical reports and additional publications is appended. (Author/JB)

14/7/40 (Item 1 from file: 7)
DIALOG(R)File 7:Social SciSearch(R)

00010014 Genuine Article#: M0237 Number of References: 4
Title: AUTOMATIC KEYWORD CLASSIFICATION **FOR INFORMATION RETRIEVAL - JONES,KS**
Author(s): **SALTON G**
Corporate Source: CORNELL UNIV,DEPT COMP SCI/ITHACA//NY/14850
Journal: JOURNAL OF DOCUMENTATION, 1972, V28, N1, P78-80
Language: ENGLISH Document Type: BOOK REVIEW